Identification and validation of transcription factors that regulate chromatin dynamics

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von
Anne Schöler
aus Deutschland

Basel, 2012

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Prof. Dr. Dirk Schübeler und Prof. Dr. Lukas Sommer

Basel, den 26.Juni 2012

Prof. Dr. Martin Spiess
Dekan

Acknowledgements

First of all, I want to thank Dirk for giving me the chance to conduct my PhD studies in his lab. He was always open for critical discussions and gave me constructive feedback whenever needed, while his passion for science inspired me. I further appreciate that he allowed me to work independently and explore my own scientific ideas.
Moreover, I want to thank the whole Schübeler lab, with its past and present members, for the productive and highly collaborative working atmosphere that made my work in the lab and generally my time in Basel very enjoyable.
Further, I would like to thank Erik van Nimwegen and Phil Arnold for a very successful and productive collaboration that never seems to come to an end.
I want to thank Michael Stadler, Robert Ivanek, Tim Roloff and Sylvia Tippmann for their patience and the many help they gave me introducing me to bioinformatics.
I thank Chrisi, my family and my friends for ever patiently listening to me talking about my work and never giving up trying to understand it and more importantly for reminding me that there are other things in life than research.
Ultimately, I would like to acknowledge Lukas Sommer, Erik van Nimwegen, Michael Stadler and Fred Meins Jr. for serving in my thesis committee, for contributing helpful ideas during the thesis committee meetings and for evaluating this thesis.

Abbreviations

bHLH	Basic helix loop helix
bp	Basepair
ChIP	Chromatin immunoprecipitation
CtBP	Co-repressor carboxy-terminal binding protein
EED	Embryonic ectoderm development
EMT	Epithelial mesenchymal transition
EZH	Enhancer of zeste homologue
HAT	Histone acetyltransferases
HDAC	Histone deacetylase
HP1	Heterocromatic protein 1
kb	Kilobase
LMRs	Low methylated regions
LSD1	Lysine-specific demethylase 1
MBD	Methyl-CpG binding domain
MeCP2	Methyl-CpG binding protein 2
MNase	Micrococcal nuclease
nt	Nucleotide
PcG	Polycomb group
PIC	Polymerase II pre-initiation complex
Pol II	Polymerase II
PRC	Polycomb repressive complex
PRE	Polycomb response elements

PSC	Posterior sex combs
PWM	Position weight matrix
qPCR	Quantitative Polymerase chain reaction
QTL	Quantitative trait loci
RbAp	Retinoblastoma-binding protein
RMCE	Recombinase-mediated cassette exchange
SNP	Single nucleotide polymorphism
SUZ	Suppressor of zeste
TBP	TATA-box binding protein
TF	Transcription factor
TFBS	Transcription factor binding site
TK	Thymidine kinase
TSS	Transcription start site

Protein names are in capital letters irrespective of species.
Gene names are in italics.

Summary

Gene expression has to be tightly regulated during all cellular processes. During embryonic development differentiating cells loose their developmental potential and acquire specific functions by activating lineage-specific genes. Gene transcription programs are regulated by transcription factors (TFs) in concert with dynamic changes in local chromatin organisation of the DNA template. Both pathways are crucial for specific reprogramming of cells. However, how TFs and chromatin marks exactly contribute to regulate gene expression programs is not fully understood. For instance, the binding patterns of most mammalian TFs are still unknown as well as how binding specificity is achieved. Chromatin modifications are highly dynamic and cell-type specific. By regulating access to the DNA template they might guide TF binding. As most chromatin modifications have simply been associated with gene activity, a central remaining question is how chromatin modifications impact on gene expression and if they are a cause or consequence of the transcriptional state of a gene.
Further it is still an open question how chromatin marks are targeted to specific loci and how they are dynamically regulated. Trimethylation of histone 3 at lysine 27 (H3K27me3) is set by the Polycomb group of proteins, which regulate body patterning during development (Schwartz and Pirrotta, 2007; Schuettengruber et al., 2007). Polycomb-mediated H3K27me3 is associated with gene repression and essential for cellular differentiation. Further work shows that H3K27me3 targets are cell-type specific and highly dynamic during differentiation (Mohn et al., 2008; Mikkelsen et al., 2007; Bracken et al., 2006). It is unclear how these changes are regulated. Thus, we hypothesise that TFs, by recognising distinct DNA motifs, could contribute to the required specificity of chromatin reprogramming. In collaboration with the group of Erik van Nimwegen we applied an unbiased approach to model changes in H3K27me3 methylation during *in vitro* neuronal differentiation in terms of predicted transcription factor binding sites. This approach predicts many TFs to regulate H3K27me3 at specific stages of cellular differentiation. We experimentally focus on the validation of the RE-1 silencing transcription factor (REST) and the family of SNAIL TFs, which are both

predicted to regulate a gain of H3K27me3 levels as stem cells differentiate to neuronal progenitor cells. We determine genome-wide binding sites of REST at these two cellular stages and show that measured binding sites of REST show a high overlap with predicted ones. Mapping H3K27me3 in stem cells and progenitor cells of wild type and REST knock out (RESTko) cells shows a specific loss of H3K27me3 at promoter-proximal REST binding sites in neuronal progenitors, validating the computational prediction. Moreover, short promoter fragments containing either REST or SNAIL binding sites are sufficient to recruit H3K27me3, whereas deletion of the respective binding sites results in a significant loss of H3K27me3.
These results suggest that TFs are important contributors in the regulation of chromatin dynamics. However, further experiments are required to test if this is a general feature of TFs or a specialised role for REST and SNAIL proteins. In this context the extension of TF binding maps is crucial, as binding preferences for only 20-30% of all TFs are known at present. Extending this list, together with further perturbation experiments, will elucidate to what extent TF binding patterns can explain both changes in chromatin state as well as transcription.

Contents

Chapter 1

Introduction

All life starts from a single cell. During embryonic development this cell continuously divides to eventually give rise to distinct cell types that form complex systems, like the human body. Differentiation of totipotent cells into specific cell types and compartmentalisation of the developing embryo have to be precisely regulated during development. Genes associated with pluripotency have to be repressed as cells loose their developmental potential and lineage-specific genes have to be activated in a temporally and spatially correct manner. The human body consists of at least 400 highly diverse cell types (Vickaryous and Hall, 2006), which is extraordinary given that the genetic information within each cell is identical. Thus, an intense degree of regulation is required to ensure robust differentiation programs. This regulation takes place at several levels. At the level of DNA sequence transcription factors can specifically recognise and bind DNA motifs to regulate gene expression. The DNA of eukaryotes is further wrapped around histones in the form of chromatin. This compacts the DNA and impacts on the ability of TFs to bind to their cognate binding sites. In addition, histones can be chemically modified. These modifications can impact on the conformation of chromatin and are further bound by a variety of proteins. Thus, the epigenetic state of a gene adds another layer of transcriptional regulation.

Together, the gene expression pattern of a given cell is likely a complex function of transcriptional and epigenetic determinants, which is underlined by the fact that both pathways are essential for cellular differentiation. As the field of epigenetics is still quite young, there are many unanswered questions as how epigenetic modifications are targeted and how they are read-out and influence transcription. Another important question is how transcriptional and epigenetic mechanisms are causally connected.

In the following chapters I will give a more detailed introduction about the basics of mam-

malian gene regulation, the importance of TFs and chromatin and introduce possible regulators of chromatin states.

1.1 Gene regulation in mammalian genomes

Animal genomes show a great variability in genome size ranging from 20 megabases up to 100 gigabases (Gregory, 2012). During evolution genome size increased with the rise of more complex organisms. It was initially assumed that the large genomes of mammals would harbour many more genes than simpler organisms with small genomes. However, with the sequencing and annotation of mammalian genomes, it became evident that this is not the case: An increase in genome size does not generally scale with an increase in gene number. Whereas early estimates for the number of human genes were ranging from 50.000 - 100.000 genes (Bird, 1995), the actual number is most likely between 25.00 - 30.000 (Baltimore, 2001; Lander et al., 2001). In terms of gene number the human genome is only slightly more complex than the worm *Caenorhabditis elegans*, which possesses about 20.000 genes (Ruvkun and Hobert, 1998). Thus, the total number of genes is most likely not the optimal measurement for organismal complexity. Claverie suggested to define biological complexity as the possible number of transcriptional states: Assuming in a simple model, that each human gene can be either on or off in a genome with N = 30.000 genes, this would allow the human genome to possess $2^N = 2^{30000}$ distinct transcriptional states. Relative to the worm *Caenorhabditis elegans* the human genome would thus be $2^{30000}/2^{20000} = 2^{10000} \sim 10^{3000}$ more complex (Claverie, 2001). This simple calculation demonstrates that even a small increase in gene number can generate a large increase in organismal complexity, implying that mammals likely possess sophisticated regulatory mechanisms to generate distinct cell types from a rather "limited" set of genes.

Transcription factors are integral regulators of all transcriptional processes. By binding to promoter-proximal and distal regulatory regions such as enhancers, TFs regulate correct spatial and temporal gene expression patterns. Enhancers are thought to act as primary determinants of tissue-specific gene expression and have been characterised in detail (Buecker and Wysocka, 2012). However, detailed knowledge regarding the mapping of enhancer-promoter interactions is still lacking. This project mostly focussed on promoter-proximal gene regulation.

Several lines of evidence suggest TF-mediated gene regulation to be a central component in the establishment of transcriptional programs and mammalian complexity: The number of TFs increases in the order yeast, nematode, fruit fly, human (Tupler et al., 2001), whereas the diversity of cell types increases in the same order (Carroll, 2001). Considering networks of TFs and the genes they regulate is likely a better measure of biological complexity than the mere number of genes (Szathmary et al., 2001). This idea is supported by the finding that an

increase in gene number is generally accompanied by a yet larger increase in the number of TFs (van Nimwegen, 2003). Yet, mammalian genomes are very large and the minor increase in gene number during evolution was accompanied by a great increase in non-coding DNA content. This is mainly due to an accumulation of transposable elements and repetitive DNA and results in a small protein-coding content of about 2 % (Elgar and Vavouri, 2008). Precise expression of genes is required, whereas transposable elements have to be kept silent at the same time.

A key difference between prokaryotic and eukaryotic cells is the packaging of the DNA. In prokaryotes the DNA can be associated with histone-like proteins in the cytosol. This interaction between DNA and proteins is unlikely to play a general repressive role in transcription resulting in a non-restrictive ground state in prokaryotes (Struhl, 1999). In eukaryotes the DNA is packed into the nucleus and importantly wrapped around nucleosomes. This compacts DNA and more importantly adds a level of basal repression as the accessibility of the DNA to binding factors is reduced and *in vitro* transcription is impeded (Knezetic and Luse, 1986). Thus, the transcriptional ground state in eukaryotes is restrictive (Struhl, 1999). These key differences between prokaryotic and eukaryotic cells are likely to reduce transcriptional noise. In prokaryotes the transcribed RNA is directly accessible to the translation machinery, whereas in eukaryotes transcription and translation are spatially separated and tightly regulated (Bird, 1995).

The genomic DNA of mammals is further characterised by global DNA methylation, which happens mostly in the context of CpG dinucleotides and is associated with gene repression. The packaging of DNA into chromatin and DNA methylation serve as independent mechanisms in vertebrates and are thought to be essential for the repression of spurious transcription (Bird, 1995).

Therefore, all gene regulation in mammals has to be considered in the context of chromatin, as local modulation of DNA accessibility is required for all DNA templated processes. Together, the regulated binding of TFs to proximal and distal regulatory elements, covalent modifications of nucleosomes and the position of a gene within the nucleus all influence its expression level (Zhou et al., 2011; Noonan, 2009). The final protein amount is further determined by many post-transcriptional regulatory steps such as mRNA decay and translational regulation (Turner, 2011).

Together, gene regulation in mammals is a complex network, that is still not fully understood. This work will mostly focus on gene regulation by transcription factors and covalent modification of nucleosomes.

1.2 Transcription factors

Transcription factors, which can specifically bind to DNA, are major regulators of transcription in all cellular processes such as development, signal transduction, immune responses and metabolism. TFs can be grouped into general TFs and activating/repressive TFs. General TFs, such as TFIID are essential for transcriptional initiation and highly conserved from yeast to human (Eisenmann et al., 1989). TFIID together with other general TFs and the Polymerase II holoenzyme make up the RNA polymerase II pre-initiation complex (PIC). Subsequent, transcription initiation, elongation and termination are subject to many quality controls and thought to be tightly regulated (Cooper, 2000). The second group of TFs is much larger in size and regulates promoter-specific transcription in a sequence-specific way. About 1400 TFs (Vaquerizas et al., 2009) exist in humans and their number has greatly increased during evolution (Tupler et al., 2001). TF-mediated gene regulation is thought to be a principal requirement for the emergence of metazoan life (Levine and Tjian, 2003) and changes in *cis*-regulatory sequences are a major contributor underlying morphological evolution (Frankel et al., 2011; Carroll, 2008). The constant expansion of the TF repertoire along the human lineage coincided with the emergence of increasing organismal complexity and enabled the development of new functions. For example the homeodomain family of TFs appeared during the emergence of a body plan in animals (Garcia-Fernandez, 2005). In addition a large group of about 13 % of human TFs are primate-specific (Vaquerizas et al., 2009).

The largest group of TFs in humans are the zinc finger TFs that make up about half of all TFs. Zinc fingers are small structural motifs that can coordinate one or more zinc ion to help stabilise their folds and they are very common in mammalian TFs. The group of zinc finger TFs has expanded at several evolutionary stages, including the emergence of vertebrates and most during the appearance of mammals and primates (Chen and Rajewsky, 2007; Vaquerizas et al., 2009). A possible explanation for this great expansion is that zinc finger TFs, by mutating amino acids that directly interact with the DNA, can easily change their binding specificity during evolution (Vaquerizas et al., 2009). TFs are still rapidly evolving in humans and are under positive selection (Bustamante et al., 2005) making them key candidates in explaining phenotypic differences between species (Wilson and Odom, 2009).

TF binding sites (TFBSs) are highly enriched in regulatory regions such as promoters and enhancers. It is however still unclear if promoter sequences or enhancer sequences act as the major determinants to regulate gene expression. Interestingly, the mapping of the TFs OCT4, NANOG and SOX2 in stem cells showed that most binding events are at distal reg-

ulatory regions and not at promoters (Young, 2011; Chen et al., 2008; Boyer et al., 2005). As much as 10% of the human genome is estimated to encode enhancer elements suggesting that enhancers are primary determinants of tissue-specific gene expression (Buecker and Wysocka, 2012). To further explore this, comprehensive mapping of enhancer-promoter interactions will be required.

The large increase in genome size during evolution generates an increase in potential TFBSs, and requires means of mammalian genomes to strictly regulate TF binding. One might think that primate-specific TFs contain longer DNA recognition motifs to ensure specific binding. However, the opposite seems to be the case. Bacterial TFs tend to have longer binding motifs than mammals. For example, the sigma factor binding site in *E. coli* has 12 conserved positions (Lisser and Margalit, 1993) whereas the analogous eukaryotic TATA box is only 6 bps long (Bucher, 1990). Differences in motif length could reflect different mechanisms to control binding specificity in prokaryotes versus eukaryotes (Bilu and Barkai, 2005). Interestingly, prokaryotic genes are typically bound by a single TF, whereas eukaryotic promoters often contain many different TF binding sites (Wray et al., 2003) arguing for combinatorial regulation of eukaryotic TFs. Most mammalian TFs recognise short motifs between 6 and 8 bps. Many of these motifs are further degenerate leading to millions of potential binding sites of which only a subset is bound *in vivo*. This strongly argues that functional and non-functional binding sites are discriminated by additional means. One possibility is that cooperative or sequential binding of TFs could generate functional binding sites. First, a "pioneering" TF would bind, possibly to a DNA stretch located between nucleosomes, which would then induce remodelling events that allow other TFs to bind (Zaret and Carroll, 2011). Recent work from the Young lab suggested that master TFs such as MYOD1 and OCT4 direct SMAD3 binding to DNA and thus determine cell-type-specific effects of TGFβ signalling (Mullen et al., 2011). Another possibility is that chromatin accessibility directs TF binding. Supporting this argument is a study showing that glucocorticoid receptor binding occurs mostly at cell-type specific accessible DNase I hypersensitive sites, implying that chromatin might discriminate functional and non-functional binding sites (John et al., 2011). This project will briefly address the question whether the quality of TFBSs regulates TF binding dynamics.

In the following sections I will give a more thorough introduction for specific TFs, one binding a rather untypical 21 bp long motif and one a more typical 6 bp sequence.

1.2.1 RE-1 silencing transcription factor

The transcriptional repressor RE-1 silencing transcription factor (REST, also called neuron-restrictive silencer factor; NRSF) is a TF of the zinc finger family. It has been the subject of many studies and represents a textbook example of a vertebrate specific TF, whose binding sites have greatly increased during evolution (Johnson et al., 2009) as discussed in section 1.2. During development REST acts as a repressor of neuronal genes. Its targets include ion-channels, neurotransmitters, growth factors and hormones, as well as proteins involved in axonal guidance and vesicle trafficking (Bruce et al., 2004). Further, REST was shown to function as both a tumour suppressor and oncogene (Westbrook et al., 2005; Majumder, 2006). Proper regulation of REST is critical, as REST over-expression causes axon path finding errors (Paquette et al., 2000).

Originally viewed as a master regulator of neuronal differentiation, by now multiple studies suggest that REST does not control the induction of neurogenesis but rather maintains the repression of inappropriate differentiation genes (Chen et al., 1998; Jones and Meech, 1999). In line with this is the phenotype of REST knock-out mice. Mice deficient for REST die at embryonic day 11.5 (Chen et al., 1998), but do neither show transformation of non-neuronal cells into neurons nor induce neurogenesis of neuronal precursors. Importantly, REST knock-out embryonic stem (ES) cells are viable and show no defects in pluripotency (Jorgensen et al., 2009; Yamada et al., 2010).

REST specifically binds to RE-1 elements in the genome, which consist of a rather long 21 bp consensus sequence. Besides this canonical motif, two half site motifs were identified corresponding directly to the separate left and right sides of the canonical motif spaced by an additional 5 to 9 bp (Johnson et al., 2007). Importantly, this variable spacing drastically increases the number of potential sites, of which only a subset are actually bound by REST. Sun et al. analysed REST target genes and found 65% of genes with a RE-1 site in introns, whereas another 28% and 7% had the RE-1 site located in the 5' and 3'-flanking regions, respectively (Sun et al., 2005).

REST possesses two repressor domains, one at the C- and one at the N-terminus and a multitude of proteins have been suggested to interact biochemically with REST. In differentiated non-neuronal cells, a CoREST-histone deacetylase (HDAC) complex has been reported to bind to the C-terminal repressor domain of REST (You et al., 2001; Andres et al., 1999) to recruit lysine-specific demethylase 1 (LSD1) (Lee et al., 2005) and histone H3K9 methyltransferase G9a (Shi et al., 2003). The additional recruitment of methyl-CpG-binding protein 2 (MeCP2) (Nan et al., 1998) and heterochromatin protein 1 (HP-1) have been suggested to induce a compact chromatin conformation. Further the N-terminal repressor domain of

REST recruits the corepressor Sin3a, HDACs (Huang et al., 1999) as well as MeCP2. It is thought that this multitude of epigenetic modifications ensures the stable repression of neuronal genes in non-neuronal tissues (Ballas and Mandel, 2005), yet most of the proposed interactions were detected at single genes and or with transient reporter systems. Which of these interactions are the most relevant at the level of the genome is still open. Interestingly, even though REST expression decreases during neuronal development it can still be detected in several regions of the adult rat brain. Further brain-specific splice variants of REST were detected at low concentrations (Palm et al., 1998). This suggests that REST functions may be more diverse than currently considered.

1.2.2 SNAIL transcription factors

The family of SNAIL TFs also belongs to the zinc-finger TF group but, unlike REST, these proteins are highly conserved during evolution (Kerner et al., 2009) and play a key role in mesoderm formation from flies to humans (Alberga et al., 1991; Nieto, 2002). Evolutionary, the SNAIL superfamily consists of two independent families *snail* and *scratch*, that grew more complex by multiple gene duplication events (Manzanares et al., 2001). In every animal that has been completely sequenced, except for *Ciona intestinalis*, at least one *snail*-like and one *scratch*-like gene can be found, showing a strong conservation of these proteins during animal evolution (Kerner et al., 2009). SNAIL TFs are thought to act as key transcriptional repressors in embryonic development, neuronal differentiation (Nakakura et al., 2001b), neural crest formation (Carl et al., 1999), cell fate decisions such as epithelial-mesenchymal transition (EMT) (Carver et al., 2001) and left-right identity (Hemavathy et al., 2000). Moreover, the expression of mouse SNAIL was shown to be tightly associated with invasive areas of squamous-cell carcinoma arguing for a role of SNAIL proteins in cancer development and progression (Batlle et al., 2000; Cano et al., 2000).

SNAIL-mediated gene repression is thought to depend on motifs found in the amino-terminal region. The SNAG (Snail/Gfi) domain is important for repression (Nakayama et al., 1998) and conserved in all vertebrate *Snail* genes (Nieto, 2002). In addition, in flies SNAIL further interacts with the co-repressor carboxy-terminal binding protein (CtBP) (Nibu et al., 1998). It was further shown that the N-terminus can recruit a Sin3A/histone deacetylase 1 complex (Peinado et al., 2004). However, the detailed mechanism of repression by SNAIL TFs has not been resolved.

In mouse the SNAIL family consists of four family members named SNAIL (Nieto et al., 1992), SLUG(Sefton et al., 1998), SMUC (Kataoka et al., 2000) and SCRATCH (Nakakura

et al., 2001a). The SNAIL protein is best studied and SNAIL-mutant mice were shown to die during gastrulation due to defective EMT (Carver et al., 2001). The *Slug* gene was shown to be neither required for mesoderm formation nor for neural crest generation and development in mice (Jiang et al., 1998). The mouse *Scratch* gene shows a neural specific expression pattern and was suggested to play a role in the regulation of neuronal differentiation (Nakakura et al., 2001a,b).
All SNAIL family members are thought to bind to a six bp CAGGTG motif (Mauhin et al., 1993). This motif corresponds to the E-box consensus sequence, which is also bound by basic helix-loop-helix (bHLH) TFs. Thus, SNAIL and bHLH TFs might compete for the same binding sites (Kataoka et al., 2000). SNAIL proteins have so far been mainly characterised by their spatial and temporal expression patterns in different species and by studying the evolution of the SNAIL family. An exception, where SNAIL function is rather well characterised is the developing *Drosophila melanogaster* embryo. Here, SNAIL, together with the TFs TWIST and DORSAL, is required for dorsoventral patterning and genome-wide binding of these three TFs has been measured, identifying new potential enhancers (Zeitlinger et al., 2007). Interestingly, SNAIL, TWIST and DORSAL show very similar binding patterns, suggesting that interaction of these proteins might facilitate functional binding (Zeitlinger et al., 2007; He et al., 2011). In mammalian cells only a small number of SNAIL targets have been identified (Peiro et al., 2006), including the well-studied SNAIL target E-CADHERIN (Batlle et al., 2000). Functional insights require further characterisation of genome-wide binding patterns of the different SNAIL family members to uncover their role in cellular function.

1.3 Chromatin

Every human cell contains 3.2 gigabases of DNA, that is about 2 metres in length if fully extended (Alberts et al., 2002). Several degrees of compaction up to a factor of 10.000 are required to reach the condensation of mitotic chromosomes (Jiang and Pugh, 2009). At the lowest level of compaction 147 bps of DNA are wrapped around nucleosomes. These consist of an octamer of histones, which are small basic proteins. Each nucleosome consists of four different histone proteins H2A, H2B, H3 and H4 that are present in two copies each (Figure 1.2). DNA and histones together are visible under the microscope as "beads on a string". The DNA is further compacted by linker histone H1 into transcriptionally inactive 30 nm fibres, which then form chromosomes (Figure 1.1). Several histone variants exist, that replace the canonical histones in specific locations or biological contexts. At active genes histone H2A

and H3 are replaced by the variants H2AZ and H3.3, respectively. This might destabilise nucleosomes and maintain a chromatin structure permissive for transcription (Bell et al., 2011).

Packaging the DNA in form of chromatin obstructs access of proteins to DNA. Subsequently, promoter regions from yeast to human are generally depleted of nucleosomes, implying that important regulatory regions can be identified by their reduced nucleosomal occupancy.

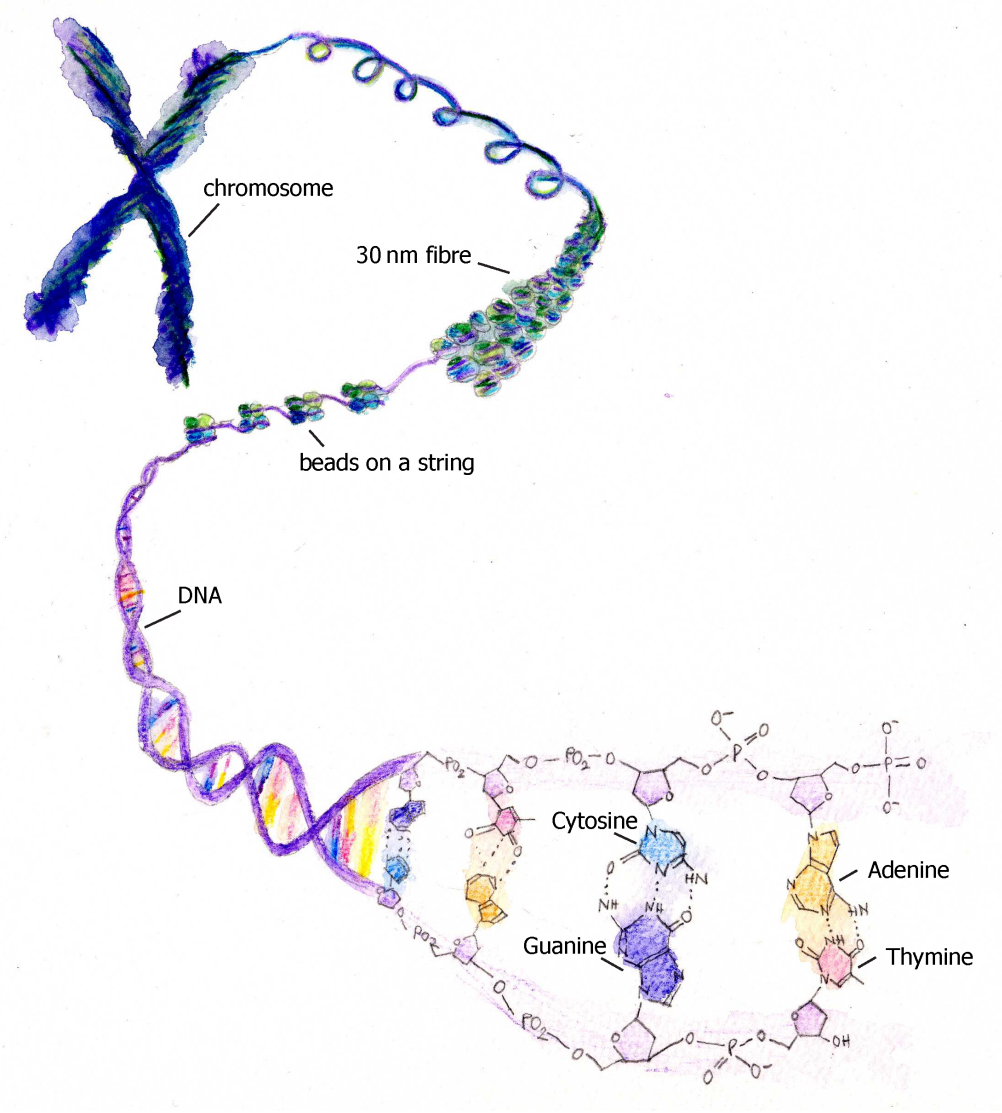

Figure 1.1: Depicted is a schematic view of the multiple compaction levels of DNA in the nucleus. At the lowest level the molecular structure of the DNA with the four bases thymine, guanine, adenine and cytosine is shown. The helical DNA is then wrapped around histones (beads on a string) and further compacted into transcriptionally inactive 30 nm fibres, which then form chromosomes.

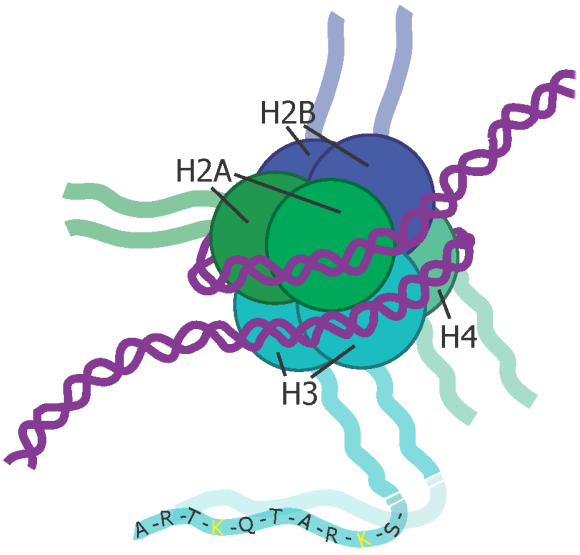

Figure 1.2: Representation of a nucleosome associated with DNA. Each nucleosome consists of two copies each of histone H2A, H2B, H3 and H4. The N-terminal histone tails protrude from the nucleosome complex. Specific residues of each histone tail (exemplified for H3), such as lysines at position 4 and 9 (K4, K9; highlighted) of H3 are subject to specific post-translational modifications (see section 1.3.1)

Indeed, multiple reports showed that TF binding can be accurately inferred from DNA sequence and chromatin accessibility data (Birney et al., 2007; Bergman et al., 2005; Pique-Regi et al., 2011). Whether these nucleosome free regions (NFRs) are generated by TF binding and/or by active remodelling of nucleosomes is not fully understood. Genome-wide analysis of nucleosomal occupancy can be carried out by Micrococcal nuclease (MNase) digestion or DNase I hypersensitivity analysis. This revealed that both DNase I hypersensitive sites and regions depleted of nucleosomes are overlapping with regulatory regions such as enhancers and promoters (Wu et al., 1979; Elgin, 1988; Birney et al., 2007). Nucleosome localisation can be altered by thermal motion, protein binding and remodelling by chromatin remodelling enzymes that can slide nucleosomes along the DNA or even evict them temporarily (Bell et al., 2011). TF binding to nucleosomal DNA *in vitro* can directly lead to displacement of nucleosomes (Workman and Kingston, 1992). All DNA templated processes, such as transcription, replication and DNA repair happen in the context of chromatin. Thus, nucleosome dynamics are crucial for proper gene regulation and transcription fidelity (Jiang and Pugh, 2009). Interestingly, NFRs are present at promoters irrespective of the transcriptional state. They are permissive for transcription but not sufficient to activate genes (Jiang and Pugh, 2009). For transcription initiation chromatin remodelling enzymes are required, such as RSC, which can both evict and reposition nucleosomes and is required for activation of many yeast genes (Parnell et al., 2008). Gene activation is tightly regulated as transcriptional initiation at cryptic start sites is prevented by another chromatin remodelling enzyme Isw2 (Whitehouse et al., 2007). The finding that *in vitro* reconstitution of nucleosome positioning outside of yeast promoters requires ATP-dependant trans-acting factors further emphasises the importance of chromatin remodelling enzymes (Zhang et al., 2011). In summary, nucleosome positioning and occupancy is likely determined by a combination of TF binding, DNA sequence features, nucleosome remodelling and histone modifiers

(Bell et al., 2011). Chromatin state might further direct enzymes to their appropriate sites of action. Chromatin modifications and proteins specifically binding to chromatin are thought to help distinguish non-coding and non-regulatory regions of DNA from regulatory regions. Moreover, it is hypothesised that chromatin state specifies functional from non-functional TFBSs. Interesting work from the Pritchard lab suggests that DNase I hypersensitive sites are formed by TF binding and can explain expression variation between individuals (Degner et al., 2012). Further work in yeast showed that during oxidative stress nucleosome eviction at the binding sites of the TF MSN2P occurred after TF binding (Huebert et al., 2012). These data argue that TFs might be able to bind their cognate sites even at nucleosome occupied regions. However, as this happens mostly at regulatory regions it is possible that TF binding is precluded by a distinct chromatin state at non-regulatory regions.

1.3.1 Modifications of histones and DNA

The highly basic histone proteins attract and neutralise the negative charge of DNA. However, the fact that histone proteins are highly conserved from yeast to humans already suggests that histones play more than a structural role. Indeed, once viewed as merely packaging material it is now evident that histone proteins are chemically modified and that these modifications are cell-type and cell-stage specific. At present up to 100 different post-translational modifications have been identified including methylation, acetylation, phosphorylation and ubiquitination (Kouzarides, 2007; Bernstein et al., 2007), which occur at specific residues of the N-terminal histone tails that protrude from the nucleosome (Luger et al., 1997)(and see Figure 1.2). Modified residues include lysines (K), arginines (R), serines (S) and threonines (T) (Kouzarides, 2007). Of all enzymes that modify chromatin the enzymes that set methylation and phosphorylation marks are the most specific (Kouzarides, 2007).
This wide array of histone modifications regulates accessibility of DNA and further allows specific interactions with effector proteins. Here, specific domains such as chromo- and tudor-domains recognise methylation, whereas bromodomains recognise acetylation (Kouzarides, 2007). Deletion of histone tails or certain residues results in specific effects on gene expression in yeast (Kayne et al., 1988; Nakanishi et al., 2008; Dai et al., 2008).
Another important epigenetic modification is DNA methylation, which in mammals occurs almost exclusively in the context of CpG dinucleotides. Most CpGs in the genome are methylated with the exception of high-density CpG regions, termed CpG Islands (Bird, 1986; Carninci et al., 2006). Regions of reduced DNA methylation are mostly found at promoters and low-methylated regions (LMRs), which frequently overlap with enhancers (Stadler et al.,

2011). Methylated DNA can be recognised by methyl-CpG binding domain (MBD) proteins, which are thought to mediate repression (Bird and Wolffe, 1999). Genome-wide promoters can be separated into CpG-rich and CpG-poor promoters (Bird, 1986; Balwierz et al., 2009). CpG-rich promoters typically have loosely defined start sites and regulate housekeeping genes (Carninci et al., 2006), whereas CpG-poor promoters contain precise start sites, rely mostly on initiation via TATA-box binding protein (TBP) and regulate many tissue-specific genes (Mohn and Schubeler, 2009; Weber et al., 2007).

The variety of histone modifications together with DNA methylation is thought to demarcate regulatory regions, while keeping repetitive and non-coding regions silent. In agreement with this the genomic DNA inside the nucleus can be cytologically separated into euchromatin and heterochromatin (Heitz, 1928). Euchromatic regions only make up about 5% of the genome, are gene-rich, accessible and transcribed and carry activating chromatin modifications. Heterochromatic regions make up most of the genome, are generally gene-poor, condensed and carry histone modifications associated with a transcriptionally inactive state (Grewal and Elgin, 2002; Bannister et al., 2001). Over the past years emerging evidence has established a central role for epigenetics in gene regulation during embryonic development (Li et al., 1992; Erhardt et al., 2003), imprinting (Paulsen and Ferguson-Smith, 2001), X-inactivation (Chang et al., 2006), and the control of transposons (Bourc'his and Bestor, 2004).

However, functional and mechanistic insights are still lacking as the direct impact on transcription and specific targeting mechanisms are still unclear for most epigenetic marks. A common view is that epigenetic modifications serve as an additional layer of gene repression to increase the robustness of differentiation programs and suppress transcriptional noise (Bird, 1995; Pujadas and Feinberg, 2012). Proposed specifiers of chromatin state include DNA sequence, DNA methylation patterns, TFs or other regulatory proteins and transcriptional activity (Zhou et al., 2011). To date, the possible function of most chromatin marks has been characterised by correlating them to genomic features such as promoters, genes, enhancers and to gene expression levels. This has revealed several histone modifications that are generally associated with an active state of transcription such as methylation at lysine 4 of histone H3 (H3K4) (Santos-Rosa et al., 2002), H3K36 (Krogan et al., 2003), H3K79 (Schubeler et al., 2004) and histone acetylation. Methylation of H3K27 (Cao et al., 2002), H3K9 (Bannister et al., 2001) and histone deacetylation (Taunton et al., 1996) correlate with a repressed state.

Epigenetic patterns are stably retained during somatic cell divisions and can be dynamically regulated during cellular differentiation (Mohn et al., 2008; Mikkelsen et al., 2007, 2010). With the exception of DNA methylation, where the mechanism of inheritance during cell cycle is well established (Law and Jacobsen, 2010), the mode of propagation for most epi-

genetic marks is not clear. Different models exist such as positive feedback loops between the mark and the enzyme that sets it. Such cooperative behaviour has been described for propagation of H3K9 and H3K27 methylation (Margueron and Reinberg, 2010; Hansen et al., 2008; Margueron et al., 2009; Nakayama et al., 2001). How chromatin marks are targeted de novo is still largely unclear and further discussed in the following sections.

1.3.2 Epigenetic marks associated with gene activation

The exact role of most histone modifications is still unclear and many modifications are characterised by their correlation to the process of transcription. Active genes are methylated at H3K4, H3K36, H3K79 and acetylated at several residues of H2A, H3 and H4 (Figure 1.3). Acetylation is historically the most studied modification. More recently, the focus shifted to lysine methylation with methylation at H3K4 and H3K36 being the best studied modifications associated with gene activity.

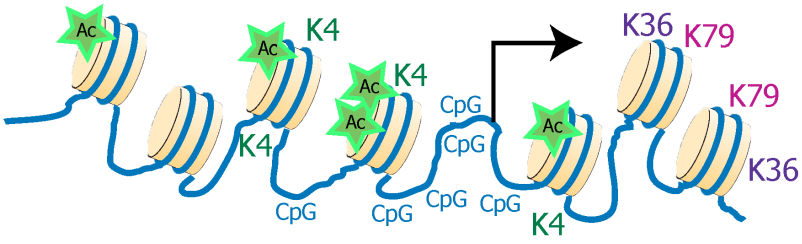

Figure 1.3: Schematic representation of chromatin modifications associated with gene activity. Promoters of active genes generally have low nucleosome occupancy, low DNA methylation (CpG). Nucleosomes around the TSS are acetylated (Ac) and methylated at H3K4 (K4). Methylation of H3K36 (K36) and H3K79 (K79) are enriched in gene bodies.

H3K4 can be mono-, di- and trimethylated. H3K4 mono-methylation (H3K4me1) is abundant downstream of the TSS and at enhancers and has been used as a criterion to define these regulatory regions (Heintzman et al., 2007; Birney et al., 2007). Di- and trimethylation of H3K4 (H3K4me2/3) are strongly enriched at CpG Islands including many promoters (Mikkelsen et al., 2007; Lee and Skalnik, 2005; Barski et al., 2007). As CpG Islands are methylated at H3K4 irrespective of their transcriptional status the methylation is not necessarily a predictor of expression at these promoters (Weber et al., 2007; Mohn et al., 2008). However, at weak or CpG-poor promoters H3K4me2 is a better predictor of transcription levels as these regions show no H3K4me2, when they are not expressed (Mohn and Schubeler,

2009). It was shown that H3K4me3, is specifically bound by bromodomain and PHD finger transcription factor (BPTF), which is part of the NURF complex. NURF is an ATP-dependent chromatin-remodelling complex that disrupts chromatin to enhance initiation of transcription (Wysocka et al., 2006). Moreover, methylated H3K4 is specifically recognised by chromatin remodelling protein Chd1 (Sims et al., 2005; Pray-Grant et al., 2005) suggesting a mechanism for H3K4 methylation in increasing chromatin accessibility and facilitating transcription initiation. H3K4me3 regions overlap with DNase I hypersensitive sites and H3.3-containing nucleosomes (Rando, 2007; Li et al., 2007; Birney et al., 2007). However surprisingly, mutations of H3K4 methyltransferases in yeast and mouse ES cells have little effects on steady state gene expression (Lenstra et al., 2011; Jiang et al., 2011). A possible targeting mechanism of H3K4 methylation was suggested by the Bird lab with the identification of Cfp1, a protein that specifically binds unmethylated CpGs and interacts with Setd1 H3K4 methyltransferase (Thomson et al., 2010).

It is still unclear why and how CpG Islands are protected from DNA methylation. As methylation of CpG Islands is generally accompanied by a loss if H3K4 methylation it was suggested that H3K4 methylation might play an active role in protecting from DNA methylation. Interestingly, methylation of H3K4 was shown to preclude physical interaction between the histone tail and DNA methyltransferase 3-like protein (Ooi et al., 2007). Several enzymes have been identified that set and remove K4 methylation, yet their specific targeting and activity is still unclear (Li et al., 2007).

Methylation of H3K36 can also occur as mono (H3K36me1), di-(H3K36me2) or trimethyl (H3K36me3) (Greer and Shi, 2012). The recruitment and function of H3K36 methylation is particularly well studied. In *Saccharomyces cerevisiae* H3K36 methylation is carried out by Set2, which is associated with elongating Polymerase II (Pol II) (Krogan et al., 2003), explaining the localisation of this modification at gene bodies. H3K36me3 is recognised by the Rpd3S histone deacetylase, which creates a hypoacetylated environment within transcribed regions that suppresses spurious intragenic transcription (Carrozza et al., 2005; Keogh et al., 2005). H3K36me3 levels over gene bodies are a very good predictor of the transcription levels of genes (Tippmann et al., MSB, *in press*). Exons show increased nucleosome density compared to introns resulting in increased levels of H3K36me3 over exons compared to introns (Schwartz et al., 2009). Differences in H3K36me3 levels across gene bodies were further suggested to regulate co-transcriptional alternative splicing possibly pointing to a role for H3K36 methylation in this process (Luco et al., 2010).

Methylation of H3K79 is catalysed by Dot1, which sets mono-, di- and trimethylation at lysine 79 of histone H3 (van Leeuwen et al., 2002). H3K79 methylation is enriched over gene bodies (Pokholok et al., 2005), yet no specific functions have yet been assigned to the

different methylation states (Frederiks et al., 2008).
Another important set of activating histone marks is acetylation of the N-terminal tails of histone H2A, H3 and H4, which are catalysed by histone acetyltransferases (HATs), which are often part of co-activator complexes (Ogryzko et al., 1996; Kuo et al., 1998; Kouzarides, 2007). Acetylation of histones influences the net charge of nucleosomes and reduces the electrostatic interaction between histones and DNA (Grunstein, 1997; Wolffe and Hayes, 1999). In line with this, acetylated histones were shown to overlap with transcribed regions (Schubeler et al., 2004; Wang et al., 2008). With the exception of H4K16ac the influence of acetylation on gene activity seems to depend less on specific lysine residues but more on the absolute level of acetylation (Dion et al., 2005). H4K16ac modulates both higher order chromatin structure, by preventing the formation of compact 30 nm fibres, and functional interactions between the chromatin remodelling enzyme ACF and the chromatin fiber (Shogren-Knaak et al., 2006). Thus, H4K16ac correlates with increased DNA accessibility at promoters and gene bodies (Bell et al., 2010). A recent study by the Kingston lab determined the structure of the *Saccharomyces cerevisiae* repressor Sir3 bromo-associated homology domain and presented structural evidence how H4K16ac might inhibit the interaction of Sir3 and H4K16 acetylated nucleosomes (Armache et al., 2011).

1.3.3 Epigenetic marks associated with gene repression

Whereas activating chromatin marks are thought to increase the accessibility of DNA for TFs the opposite is assumed for epigenetic modifications associated with gene repression. Important repressive pathways are DNA methylation, Polycomb-mediated gene repression, histone deacetylation and methylation of H3K9 and H4K20 (Taunton et al., 1996; Bannister et al., 2001; Lu et al., 2008) (Figure 1.4).

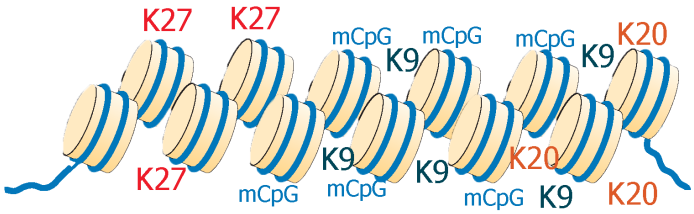

Figure 1.4: Schematic representation of chromatin modifications associated with gene repression. Repressed regions typically have high nucleosome occupancy, carry DNA methylation (mCpG) and nucleosomes are methylated at H3K9 (K9), H3K20 (K20) and Polycomb-mediated H3K27 (K27).

Most CpGs in mammalian genomes are methylated at the 5' position of the cytosine. DNA methylation is required for heritable silencing of retrotransposons and imprinted genes (Chang et al., 2006; Bourc'his and Bestor, 2004). Methylation of CpG Islands is strongly associated with gene repression of the associated gene (Mohn et al., 2008; Bird and Wolffe, 1999; Bird, 2002; Weber et al., 2007). Bisulfite sequencing can determine the methylation state of single CpGs at base pair resolution. Therefore, DNA methylation can be easily quantified, which is not the case for chromatin modifications that are measured by chromatin immunoprecipitation (ChIP) (see section 1.3.5). Genome-wide bisulfite sequencing revealed that DNA methylation is cell-type specific and dynamic in particular at distal regulatory regions that overlap enhancers (Mohn et al., 2008; Hodges et al., 2011; Stadler et al., 2011). However, the precise read-out of DNA methylation is still unclear. A variety of methyl-CpG binding proteins exist, which are thought to specifically bind the methylated cytosines and recruit chromatin modifiers such as histone deacetylases (HDACs) that induce repression (Bird, 2002).

Polycomb group proteins were originally described in *Drosophila melanogaster* as crucial regulators of body patterning (Schwartz and Pirrotta, 2007). Since then the Polycomb system has been identified as a mediator of repression of many developmental genes during cellular differentiation. The two Polycomb repressive complexes (PRC) 1 and 2 catalyse ubiquitination of H2AK119 and trimethylation of H3K27, respectively. Both histone modifications are associated with gene repression (Cao and Zhang, 2004; Wang et al., 2004). A key question of this project is to find regulators involved in targeting of H3K27me3 thus, I will give a more thorough introduction to Polycomb-mediated repression in section 1.3.3.

Opposing the activating effect of histone acetylation (as discussed in section 1.3.2) it has been comprehensively shown that HDACs can mediate the removal of acetyl-groups to compact chromatin and confer transcriptional repression (Taunton et al., 1996; Hassig et al., 1997; Alland et al., 1997). A multitude of transcriptional repressor complexes interact with HDACs to regulate chromatin accessibility (Bird and Wolffe, 1999).

As discussed in section 1.3.1 the majority of the genomic DNA consists of heterochromatin, which typically shows methylation of H3K9 and H4K20 as well as hypoacetylation of histones (Schotta et al., 2004; Grewal and Elgin, 2002). 50 % of the mouse chromosome 19 is modified with H3K9me2, indicating that K9 methylation covers large genomic regions (Lienert et al., 2011a). Methylation of H3K9 is specifically recognised by HP1, which can oligomerise to bridge nearby nucleosomes. The resulting condensation of chromatin could reduce DNA accessibility (Bannister et al., 2001; Lachner et al., 2001; Nakayama et al., 2001). The establishment of heterochromatin in fission yeast was shown to depend on the production of non-coding RNAs (Verdel et al., 2004; Buhler et al., 2006).

Polycomb group of proteins

Polycomb-meditated gene repression was first discovered and genetically defined in *Drosophila melanogaster* as a system that controls homeobox (Hox) gene expression to ensure correct body patterning (Schwartz and Pirrotta, 2008; Lewis, 1978; Schwartz and Pirrotta, 2007; Schuettengruber et al., 2007). Since then Polycomb-mediated gene regulation has been established as a highly relevant gene repression systems during development by regulating mitotic inheritance of lineage-specific gene expression patterns (Ringrose and Paro, 2004, 2007). In embryonic stem (ES) cells Polycomb proteins were shown to be crucial for self-renewal, pluripotency and reprogramming (Boyer et al., 2006; Lee et al., 2006; Pereira et al., 2010). Polycomb targets include important developmental regulators (Boyer et al., 2006) and are in part cell-type specific (Mohn et al., 2008; Mikkelsen et al., 2007; Bracken et al., 2006). Multiple studies reported misregulation of Polycomb group (PcG) proteins in cancer (Squazzo et al., 2006; Varambally et al., 2002; Sparmann and van Lohuizen, 2006), underscoring the importance of PcGs in regulating cellular identity.

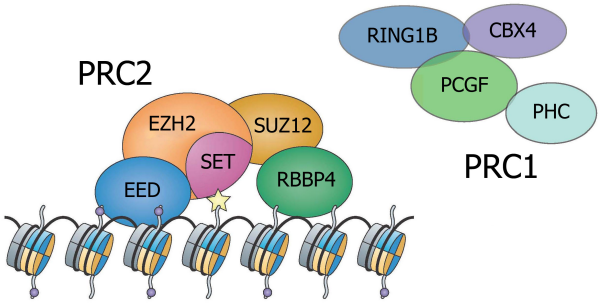

Figure 1.5: Overview of the core components of Polycomb repressive complexes 1 and 2. PRC2 consists of EZH2, SUZ12, EED and RBBP4. EZH2 contains a SET-domain that catalyses the methylation of H3K27me3. PRC1 has four members named RING1B, CBX4, PHC1 and PCGF1-6. CBX4 recognises the H3K27me3 mark and RING1B mono-ubiquitinates H2AK119. Adapted from (Margueron and Reinberg, 2010).

The Polycomb machinery consists of two multiprotein complexes named Polycomb repressor complexes (PRC) 1 and 2. In mammals PRC2 consists of four core members named enhancer of zeste homologues 1/2 (EZH1/EZH2), suppressor of zeste 12 (SUZ12), embryonic ectoderm development (EED) and retinoblastoma-binding protein p4 (RBBP4) (Figure 1.5), but multiple proteins that interact with PRC2 have been identified recently (Pasini et al., 2010; Kim et al., 2009; Tavares et al., 2012; Gao et al., 2012). Trimethylation of lysine 27 of histone H3 (H3K27me3), which is considered the hallmark of Polycomb-mediated repression is catalysed by the SET-domain-containing EZH1 and EZH2 in mammals (Czermin et al., 2002). SUZ12 is the only protein of the PRC2 complex that has a DNA binding domain (Schwartz and Pirrotta, 2007) and EED and RBBP4 are WD40-repeat-containing proteins that play a structural role. EED is essential for a functional PRC2 complex and can

specifically bind the H3K27me3 mark, suggesting a role for EED in propagation of H3K27 trimethylation (Margueron et al., 2009; Hansen et al., 2008). RBBP4 is a histone chaperone that binds to histone H4 (Verreault et al., 1996; Murzina et al., 2008). Deficiency of either EZH2, EED or SUZ12 results in early embryonic lethality in mice (Faust et al., 1995; O'Carroll et al., 2001; Pasini et al., 2004).

The PRC1 complex has four core components named RING finger containing RING1A/RING1B, chromobox protein homologue 4,6,7,8 (CBX4,6,7,8), Polyhomeotic-like 1-3 (PHC1-3) and Polycomb group ring finger 1-6 (PCGF1-6) (Beisel and Paro, 2011; Schwartz and Pirrotta, 2007) (Figure 1.5). RING1A and B function as E3 ubiquitin ligases and mono-ubiquitinate lysine 119 of histone H2A (H2AK119ub). The presence of PCGF4 (also known as BMI1) enhances the catalytic activity of RING1A (Buchwald et al., 2006). The chromodomain of CBX proteins recognises the H3K27me3 mark (Fischle et al., 2003). However, binding is not necessarily specific to H3K27me3 as different binding preferences were detected for different CBX proteins (Bernstein et al., 2006b). Recent studies showed that distinct PRC1 complexes with specific subunits exist (Tavares et al., 2012; Gao et al., 2012). However, the exact function of these is still unclear. Tavares et al., for instance, showed that the PRC1 components RYBP and CBX7 are mutually exclusive and further suggested that RYBP-containing PRC1 complexes can bind to DNA in a H3K27me3-independent fashion (Tavares et al., 2012).

While the relevance of Polycomb-mediated repression is clearly established, two major questions regarding the mode of repression and targeting to specific genes are still unclear and subject of intense research. Several studies suggested mechanisms how Polycomb-binding would mediate repression. The dogma is a step-wise process, where the PRC2 complex trimethylates H3K27, which is recognised by CBX proteins that are part of the PRC1 complex. Subsequently, RING1 proteins ubiquitinate H2AK119, which results in transcriptional repression (Wang et al., 2004; de Napoles et al., 2004). This dogma however was recently challenged by several findings. Multiple studies suggested that PRC1 can bind to genomic regions independent of PRC2 (Schoeftner et al., 2006; Tavares et al., 2012). Regarding Polycomb-mediated gene repression it was proposed that H2AK119ub blocks transcriptional elongation (Stock et al., 2007; Brookes et al., 2012), however PRC1 complexes lacking ubiquitination activity can still silence target genes (Eskeland et al., 2010). Thus, further studies are required to uncover the role of H2AK119ub in repression.

Moreover, studies suggest that Polycomb components can promote compaction of nucleosomes *in vitro* (Francis et al., 2004) and can mediate long-range interactions *in vivo*, implicating that Polycomb proteins might establish repression via higher-order chromatin structures (Tiwari et al., 2008; Lanzuolo et al., 2007; Noordermeer et al., 2011). In line with these

findings genomic methylation footprinting in *Drosophila melanogaster* revealed reduced accessibility of methylase activity at H3K27me3 domains (Bell et al., 2010). Based on recent work in Drosophila (Enderle et al., 2010) and mouse stem cells (Landeira et al., 2010) it has been suggested that Polycomb might repress by stalling polymerases.
This project does however not address the function of Polycomb proteins but rather their targeting to specific loci in the genome. Despite multiple lines of evidence for sequence-specific recruitment, Polycomb sites can neither be efficiently predicted in mammals nor in flies. Proposed targeting mechanisms are discussed in the following section.

Proposed targeting mechanism of Polycomb

As discussed in the previous section Polycomb binding is highly dynamic and cell-type specific. In addition multiple studies showed that Polycomb binding is misregulated in cancer (Squazzo et al., 2006; Richly et al., 2011). Thus, the question of how Polycomb binding is regulated and targeted to specific loci in the genome has been the subject of intense study over the past years. As Polycomb binding occurs frequently at promoters, where it is associated with a repressed state of gene expression, most work has focused on promoter-proximal Polycomb binding. Promoter-distal Polycomb regions have been studied to a lesser extent.
DNA sequence, protein-protein interactions as well as RNA have all been implicated in Polycomb targeting. That DNA sequence might be sufficient to recruit Polycomb has been a long standing dogma as Polycomb response elements (PRE) in *Drosophila melanogaster* are strongly enriched in TF binding sites. However, PREs are not defined by a consensus sequence, they rather contain many conserved motifs (Horard et al., 2000). In addition Polycomb bound regions in mammals, are often overlapping CpG islands indicating that the CpG content of promoters is a major predictor of Polycomb targeting (Ku et al., 2008; Mendenhall et al., 2010; Mohn et al., 2008; Mohn and Schubeler, 2009). Mendenhall et al. suggested that CpG rich regions depleted of activating motifs are sufficient to bind Polycomb. Another recent study showed that a tested CpG-rich sequence was sufficient for Polycomb recruitment in vertebrates (Lynch et al., 2011).
Multiple proteins have been reported to interact with Polycomb and direct specific targeting. JARID2 is a component of PRC2 in ES cells and was reported to regulate Polycomb-targeting (Pasini et al., 2010; Peng et al., 2009; Li et al., 2010). However, *in vitro* biochemical studies suggest that JARID2 is a promiscuous DNA-binding protein without particular specificity for GC-rich sequences (Kim et al., 2003; Zhou et al., 2011). Further the adipocyte enhancer-binding protein 2 (AEBP2) was shown to interact with PRC2 and proposed to play a role

in targeting of Polycomb (Kim et al., 2009). Another recent paper suggested core TFs to be involved in targeting of PRC1 (Yu et al., 2012), further arguing for a PRC2-independent role of PRC1. It was moreover shown that histone deacetylation of H3K27 by NuRD specifies local PRC2 recruitment and methylation of H3K27 in ES cells (Reynolds et al., 2011). In addition, the protein Polycomb-like 3 was recently shown to be a component of PRC2 and suggested to promote PRC2 binding to CpG Islands (Hunkapiller et al., 2012). Yet, another proposed mechanism is that Polymerase stalling might recruit Polycomb (Stock et al., 2007; Brookes et al., 2012) potentially by producing short RNAs that are transcribed from Polycomb targets and interact with PRC2 (Kanhere et al., 2010).
Several non-coding RNAs have been implicated in the targeting of Polycomb (Rinn et al., 2007; Gupta et al., 2010). For instance, short non-coding RNAs interact with Polycomb and are involved in X-inactivation in female mammals (Zhao et al., 2008; Wang et al., 2001). The non-coding RNA Hotair was suggested to act as a scaffold and a local determinant of Polycomb targeting. It is plausible that RNAs can act as important structural components of protein complexes, which is also the case for the Rox2 RNA in flies. It is however less clear how a single RNA could act as a specifier of binding to distinct regions in the genome. As TFs can attribute the required specificity to Polycomb targeting this project focused on the unbiased prediction and validation of such candidate TFs (see section 1.3.6).

1.3.4 Dynamics of chromatin states

Genome-wide mapping of chromatin states has shown that chromatin modifications are highly dynamic during cellular differentiation and reflect specific cell types (Mohn et al., 2008; Mikkelsen et al., 2007; Zhou et al., 2011; Hirabayashi and Gotoh, 2010). Most of the attention focused on dynamics at regulatory elements such as promoters, gene bodies and enhancers. Multiple studies have been conducted measuring chromatin states in stem cells versus differentiated cell types such as neurons or fibroblasts (Mohn et al., 2008; Mikkelsen et al., 2007) or addressing changes that happen during EMT or adipogenesis (McDonald et al., 2011; Mikkelsen et al., 2010). Global changes in chromatin state can be correlated to changes in gene expression and allow functional implications of chromatin dynamics.
This, showed that de novo methylation of CpG-poor regions during cellular differentiation is generally accompanied by a loss of Pol II and H3K4me2, leading to the conclusion that DNA methylation induces stable gene repression (Meissner, 2010; Mohn et al., 2008; Weber et al., 2007). More generally, de novo methylation of CpG-rich promoters and distal regions was frequently observed, whereas almost no demethylation events were detected during differen-

tiation (Mohn et al., 2008; Meissner, 2010). This suggests that DNA-methylation-mediated repression increases during lineage-specification.

The measurement of genome-wide dynamics of DNA methylation during neuronal differentiation has revealed frequent changes of DNA methylation at LMRs (Stadler et al., 2011). LMRs in ES cells are highly enriched for pluripotency factors and are frequently de novo methylated in neuronal progenitor cells. Conversely, novel LMRs appear that are enriched in neuronal specific TFs. This suggests that TF binding shapes the methlyation state at LMRs and has important implications regarding the interplay of TF binding and DNA methylation as well as the formation of unmethylated regions (Stadler et al., 2011).

Several studies showed that many genes involved in neuronal development are targeted by Polycomb in ES cells (Mohn et al., 2008; Boyer et al., 2006; Pan et al., 2007). Many of these are activated upon neuronal differentiation. However, the loss of Polycomb at these target genes in neuronal progenitors is accompanied by a gain of Polycomb at other genes many of them having specific neuronal functions (Mohn et al., 2008). This shows that upon neuronal differentiation many genes that will only be activated in terminally differentiated neurons become transiently bound by Polycomb at the intermediate progenitor stage. Moreover, a role for the H3K27me3 demethylases JMJD3 and UTX in differentiation was shown as JMJD3 is required for neuronal commitment (Burgold et al., 2008; Jepsen et al., 2007) and UTX was shown to be recruited to heart-specific enhancers, where it regulates the switch of these enhancers to an active state (Lee et al., 2011). These data further suggest that dynamic Polycomb targeting is required for cellular differentiation. Interestingly, promoters that are bound by Polycomb in ES cells where further shown to be more likely to be de novo DNA methylated upon differentiation compared to promoters that are not bound by Polycomb (Mohn et al., 2008).

Another recent study measured epigenetic dynamics during EMT and detected a global decrease in H3K9 and an increase in H3K4 and H3K36 methylation (McDonald et al., 2011). Several studies focused on the chromatin dynamics at enhances, which were shown to be much more variable and cell-type-specific than chromatin patterns at promoters (Heintzman et al., 2009; Buecker and Wysocka, 2012; Hawkins et al., 2011). A key limitation however is the lack of comprehensive mapping of enhancer-promoter interactions.

Together, these studies establish that chromatin states are cell-type specific and show that focusing on chromatin dynamics during cellular differentiation can reveal regulatory principles.

1.3.5 Crosstalk between histone modifications

The interdependency between histone modifications and DNA methylation is still not fully understood. Some histone modifications were shown to be dependent on upstream chromatin modifying events such as ubiquitination of H2BK123 was shown to be required for methylation of H3K4 and H3K79 (Briggs et al., 2002). Another study found that phosphorylation of H3S10 reduces binding of HP1 to methylated H3K9 (Fischle et al., 2005) suggesting a regulatory mechanism how these two epigenetic pathways interact. Gehani and coworkers showed that phosphorylation of serine 28 at already trimethlylated H3K27 leads to displacement of PcG proteins and subsequent gene activation (Gehani et al., 2010) suggesting a mechanism how Polycomb mediated repression could be resolved. Together with structural data these results can reveal general regulatory principles and consequences of histone modifications.
Multiple studies detected genome-wide correlations or anti-correlations of given histone modifications such as the anti-correlation of Pol II and H3K27me3 (Mohn et al., 2008; Bracken et al., 2006; Pan et al., 2007), the mutual exclusive behaviour of H3K27me3 and H3K9me2 (Lienert et al., 2011a; O'Geen et al., 2007) and the co-occurance of H3K27me3 and H3K4me3 at "bivalent" promoters in stem cells (Bernstein et al., 2006a). Yet, these results have not generated functional or mechanistic insights. One limitation is that comparison of ChIP-seq data is not quantitative. The enrichment of a given modification strongly depends on antibody quality and does not relate to absolute levels. The co-occurance of two modifications on the same nucleosome has so far only been shown for single loci, as there is a lack for methodologies that allow genome-wide quantitative mapping of histone modifications in single cells. Thus, without quantitative information it is difficult to address the role of crosstalk between histone modifications. A recent study indicated that methylation of H3K4 or H3K36 inhibits the activity of PRC2 to methylate H3K27 *in vitro* (Schmitges et al., 2011). This has several implications: Active chromatin modifications might serve as boundary elements to prevent the spreading of H3K27me3, while H3K27me3 would have to be deposited prior to methylation of H3K4 to generate bivalent regions (Schmitges et al., 2011).
Regarding the possible interplay of histone modifications the hypothesis of a histone code, where "distinct histone modifications, on one or more tails, act sequentially or in combination to form a histone code that is, read by other proteins to bring about distinct downstream events" (Strahl and Allis, 2000) has been actively debated. However, several genome-wide studies suggested that histone modifications occur in few independent combinations, implying that these patterns are more likely the result, rather than the cause, of transcription (Liu et al., 2005; Schubeler et al., 2004; Rando, 2012). Studies in yeast, where single residues of histone proteins can be easily mutated are a powerful tool in the study of interplay of

histone modifications. Dai and coworkers generated a large library of H3 and H4 mutants and analysed the impact of specific mutations on cell viability, chemical sensitivity and transcriptional silencing (Dai et al., 2008). This analysis identified several residues of H4 that are required for H3K79 methylation.

1.3.6 Inference of transcription factors that regulate chromatin states

As eluded in the previous sections chromatin dynamics have been intensively studied over the past years establishing a central role for epigenetics in gene regulation and cellular differentiation. Thus, a key question is to understand how chromatin dynamics are regulated. We focus on H3K27me3 targeting, which is set by the Polycomb system, a regulator of gene repression (see section 1.3.3). Under the premise that TFs act as local determinants of targeting the aim of this project is the unbiased identification and validation of candidate TFs that are involved in targeting chromatin marks, in particular H3K27me3.

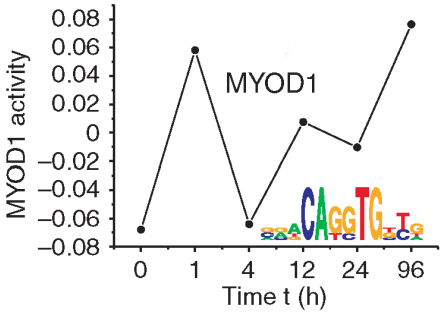

Figure 1.6: MARA predicts TFs that explain changes in mRNA levels during growth arrest and differentiation of THP-1 cells. Depicted is the inferred motif activity for the TF MYOD1 over a time course of 96 h. Adapted from (Suzuki et al., 2009)

Computational approaches that identify motifs that are overrepresented in Polycomb-bound regions found many motifs to be significantly enriched (Liu et al., 2010). However, such analyses strongly depend on search parameters and typically predict many motifs that are enriched at a single-state, making experimental validation difficult. For this project, we therefore focused on the dynamic binding of Polycomb during *in vitro* neuronal differentiation. To infer canditate TFs that regulate Polycomb dynamics in an unbiased and sophisticated manner we collaborated with the computational modelling group of Erik van Nimwegen. The van Nimwegen group implemented an approach based on genome-wide annotations of mammalian TSS that have comprehensively identified promoter regions (Harbers and Carninci, 2005; de Hoon and Hayashizaki, 2008; Balwierz et al., 2009). Using collections of regulatory motifs (Wasserman and Sandelin, 2004) and comparative genomic methods (van Nimwegen, 2007) they predicted TFBSs in proximal promoters regions genome-wide. Such resources have already successfully been used to ask to what extent TFBSs can explain patterns of gene expression (Beer and Tavazoie, 2004; Gao et al., 2004; Das et al., 2006; Suzuki et al., 2009).

In this context, the van Nimwegen group recently developed an approach termed Motif Activity Response Analysis (MARA) that identifies TF motifs that explain changes in mRNA expression. With extensive validation experiments they showed that MARA can reconstruct core transcription regulatory networks in human cells *ab initio* (Suzuki et al., 2009). MARA first predicts TFBSs at promoters genome-wide and than models changes in gene expression in terms of predicted sites. As output MARA shows for each TF motif its predicted activity. Figure 1.6 shows as an example the motif activity of the TF MYO1D during a time course of growth arrest and differentiation of leukemia cells (Suzuki et al., 2009). A positive motif activity implies that the promoters that are bound by MYO1D are expressed at that particular stage, whereas a negative motif activity predicts that the bound promoters are not expressed. The power of MARA stems from the fact that it models *changes* in gene expression. We therefore extended this approach to ask to what extent dynamic changes in chromatin can be explained by local TFBS occurrence, with the aim of identifying TFs that regulate these changes.

Chapter 2

Scope of this thesis

Cellular differentiation entails organised changes in gene expression. Pluripotent stem cells that commit to a somatic fate have to stably repress pluripotency genes and activate lineage-specific genes in a temporally correct fashion. This regulation is coordinated by TFs in concert with dynamic changes in local chromatin organisation of the DNA template. These changes have recently been documented in genome-wide analyses of histone modifications and DNA methylation. Together with genetic studies epigenome maps have helped to establish the relevance of differentiation specific reprogramming of chromatin. A key question remains regarding how chromatin modifications are targeted to specific loci during differentiation.
Thus, the aim of this project is the identification and validation of transcription factors that are involved in targeting of chromatin modifications. We focus on dynamics of H3K27me3, a chromatin modification set by the Polycomb system, arguably the most relevant gene repression system during development (Simon and Kingston, 2009; Beisel and Paro, 2011; Schuettengruber and Cavalli, 2009). Polycomb targets include developmental genes in ES cells (Boyer et al., 2006) and are in part cell-type specific (Mohn et al., 2008; Mikkelsen et al., 2007; Bracken et al., 2006). Although DNA-binding factors with limited sequence specificity have been implicated in targeting of the Polycomb system in flies (Schwartz and Pirrotta, 2008; Ringrose and Paro, 2007), the question of how Polycomb targets are specified remains currently unresolved, especially in vertebrates (Simon and Kingston, 2009; Beisel and Paro, 2011).
We aim to predict and experimentally validate TFs that regulate Polycomb dynamics during differentiation. As a system of cellular differentiation we use an *in vitro* system of neuronal differentiation. Mouse embryonic stem cells are differentiated to terminally differentiated glutamatergic neurons, via a restricted multi-potent progenitor stage (Bibel et al., 2004,

2007). Mapping H3K27me3 dynamics at each stage in wild type cells and knock-out cells of TFs predicted to regulate H3K27me3 levels will test the prediction. In addition promoter regions containing either intact or mutated TFBSs can be integrated into a distinct site in the genome via Cre recombinase-mediated exchange (Lienert et al., 2011b; Feng et al., 1999). Subsequently, TF binding and H3K27me3 levels will be measured to test if the inserted sequence is sufficient to recruit Polycomb and if this recruitment depends on functional TFBSs.

Chapter 3

Results

3.1 Predicting TFs that mediate Polycomb targeting

To systematically identify transcription factors that regulate chromatin dynamics we collaborated with Erik van Nimwegen and Phil Arnold, a graduate student in Erik's group, who extended their published MARA approach (Suzuki et al., 2009), to model genome-wide patterns of epigenetic marks, and termed this approach Epi-MARA. Epi-MARA is a linear model that explains dynamics in chromatin state as a function of TFBSs.
Concretely, if M_{ps} quantifies the amount of a particular epigenetic mark M at promoter p in sample s, and N_{pm} denotes the total number of predicted binding sites for regulatory motif m in promoter p, then we assume a linear model of the following form:

$$M_{ps} = noise + c_p + \sum_m (N_{pm} * A_{ms}) \tag{3.1}$$

where c_p is the basal level of the chromatin mark at promoter p, and A_{ms} is the unknown activity of motif m in sample s, which is inferred by Epi-MARA (see Methods Section 4.1). Abstractly speaking, the activity A_{ms} quantifies how much each occurrence of motif m contributes to the level of epigenetic mark M in sample s. One can think of A_{ms} as reflecting the occupancy of TF binding at sites of motif m and the resulting effect on chromatin mark M.
Thus, whenever Epi-MARA infers a highly positive activity A_{ms}, this predicts that the binding TF recruits the chromatin mark, whereas a highly negative A_{ms} implies that the binding TF inhibits deposition of the mark. It is important to point out that it is not the aim of Epi-MARA to provide accurate fits of epigenetic profiles at individual promoters even though it models the dynamics of epigenetic marks through equation 3.1. Since the actual levels of a chromatin mark at any promoter are likely a complex function of many variables acting both *in cis* and *in trans* the simple linear model of equation 3.1 typically captures only part of the variance in epigenetic mark levels. Importantly, however the motif activities are inferred from the combined statistics of the hundreds to thousands of promoters that contain a given motif. Thus, the linear model applied by Epi-MARA effectively averages out the complications at individual promoters, and the remaining signal provides a robust statistical average activity for each motif, enabling reliable prediction of the TFs involved in chromatin mark dynamics.
To test this approach we focused on cell-type specific targeting of Polycomb-mediated H3K27 methylation. As a biological model of dynamic changes of transcriptome and epigenome we used a well characterised mouse differentiation system, which progresses from embryonic stem (ES) cells to terminal neurons (TN) through a defined neuronal progenitor state (NP) (Bibel et al., 2007, 2004; Plachta et al., 2004). We applied Epi-MARA to a dataset of H3K27me3

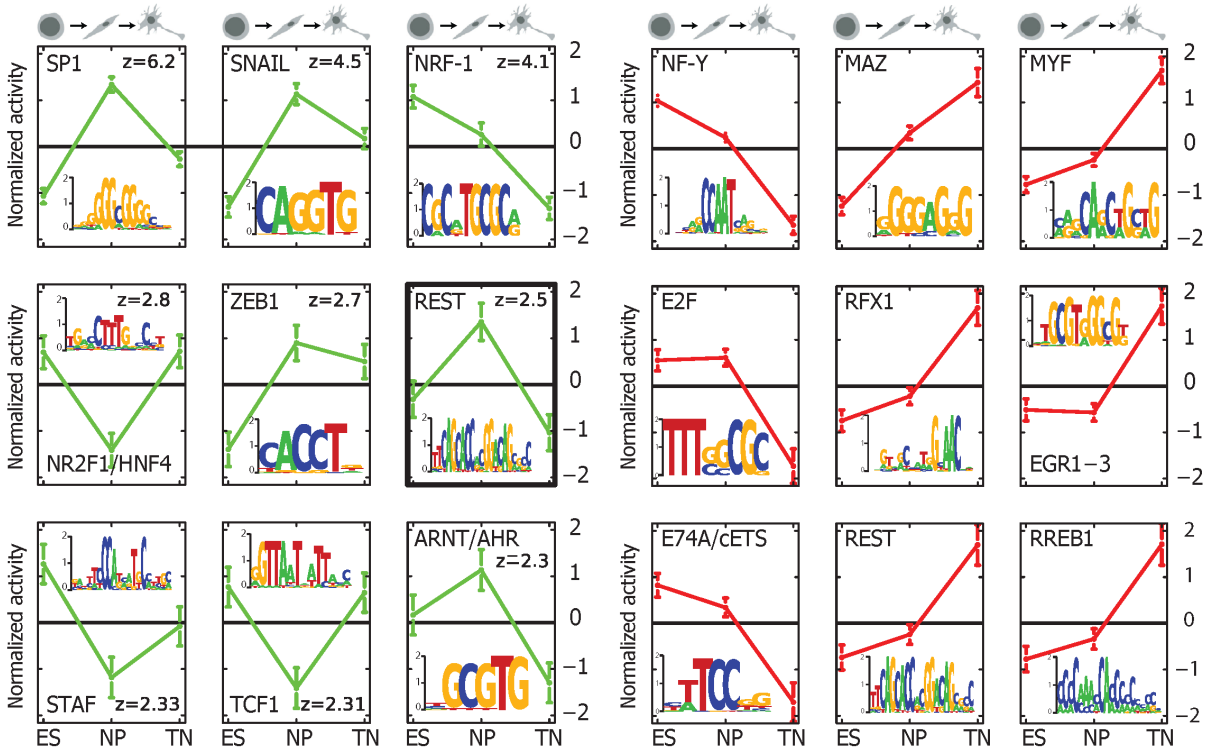

Figure 3.1: **Left**) Epi-MARA predicts transcription factors that explain dynamics in H3K27me3 levels during neuronal differentiation. Depicted are the normalised activity profiles of the top nine motifs (green lines, with standard errors indicated) with their respective z-values. **Right**) MARA expression analysis predicts the activities of transcription factors driving gene expression dynamics during neuronal differentiation: Normalised activity profiles of the nine most significant motifs that explain changes in gene expression during the differentiation process (red lines, with standard errors indicated). For both figures time points correspond to the embryonic stem cell (ES), neuronal progenitor (NP), and terminal neuron (TN) stage. Sequence logos of each of the motifs and the transcription factors thought to bind to them are shown as insets.

at promoters in the ES, NP and TN stages in the same *in vitro* neurogenesis system (Mohn et al., 2008). Figure 3.1 left shows the predicted activities of the nine motifs that contributed most to explaining the genome-wide H3K27me3 dynamics at promoters.

Five of these nine, i.e. Sp1, Snail, Zeb1, Rest, and Arnt/Ahr, show a pattern in which there is a strong transient increase in motif activity at the NP stage. That is, Epi-MARA predicts the TFs binding these motifs to be involved in the recruitment of H3K27me3 going from the ES to NP stage. Of these candidate TFs REST is of particular interest for several reasons: It is the only of the five motifs that is likely bound by a single TF and thus highly suitable for functional testing by genetic deletion. In contrast, Snail, Zeb1, and Sp1 motifs can each be recognised by multiple TFs (Nieto, 2002; Postigo and Dean, 2000; Bouwman and Philipsen, 2002), which would make rigorous experimental testing much more demanding.

Further high-quality genome-wide binding maps of REST in T cells were already available at the start of this project indicating that biological validation of predicted REST binding sites is feasible (Johnson et al., 2007).

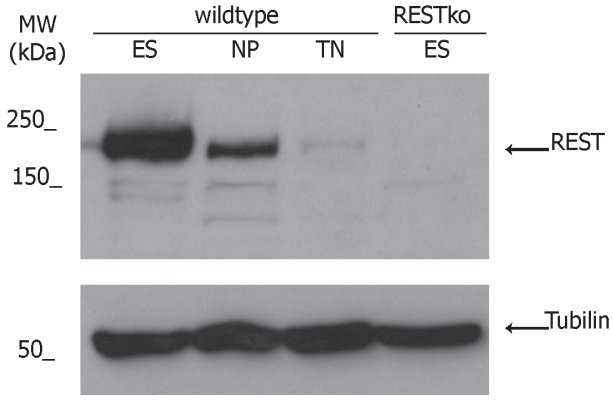

Figure 3.2: REST protein levels decrease during *in vitro* neuronal differentiation. REST protein was detected by Western Blot in extracts from the ES, NP and TN stages in wildtype as well as RESTko background (upper panel). Tubulin serves as a loading control (lower panel).

To compare the activity of TFs in regulating chromatin dynamics with their activities regulating expression we also analysed transcriptome data of the three consecutive stages using the MARA method (Suzuki et al., 2009), i.e. modelling gene expression changes in terms of TFBSs. One of the motifs that, according to the MARA analysis, most significantly regulates expression changes is E2F (Figure 3.1 right). Its inferred transcriptional activity is highly positive in the ES and NP stages where cells are proliferating, while it strongly decreases at the TN stage where cells are postmitotic and have exited the cell cycle. This is consistent with the known function of the E2F family of cell-cycle regulators that bind to this motif (Tao et al., 1997). In contrast Epi-MARA predicts no significant activity on H3K27me3 dynamics for E2F. Interestingly, the TF REST is also inferred to have an important role in driving expression changes, and its activity profile is consistent with its known role as a repressor of neuronal genes in non-neuronal tissues. As discussed in section 1.2.1 REST target genes become active at the TN stage, whereas REST itself is down-regulated (Figure 3.2). The activity profile of REST directing expression changes (Figure 3.1 right) is clearly distinct from its activity profile directing H3K27me3 (Figure 3.1 left), suggesting that REST's effects

Stage	**z-value for H3K27me3 enrichment**
Embryonic stem cell	4.2
Neural progenitor	5.7
Terminal neuron	1.3

Table 3.1: Shown is the z-value for the difference in H3K27me3 levels between predicted REST target promoters and non-target promoters. The positive numbers at all three stages indicate that REST target promoters show more H3K27me3 than non-target promoters at all stages.

on transcription levels are at least partially independent from its effects on H3K27me3 levels. Notably, the predicted cell-type specific activity for REST crucially depends on the analysis of relative changes in H3K27me3 levels across the time course as opposed to an analysis of co-occurrence of REST and H3K27me3 at individual time points. Indeed, as has been observed previously (Zheng et al., 2009; Liu et al., 2010), we find that predicted REST sites have higher than average H3K27me3 levels at all three time points (Table 3.1).

3.2 Actual and predicted REST binding overlaps

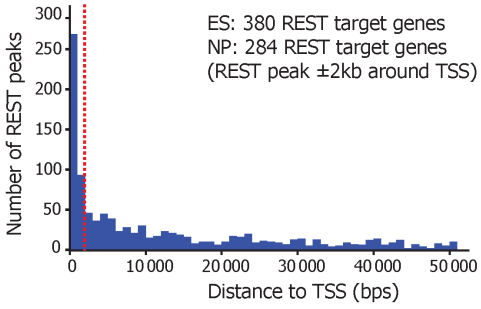

Figure 3.3: Distribution of the distance between REST binding peaks and the nearest TSS. Genes with REST binding +/- 2 kb of the TSS were classified as REST targets (cut-off indicated by dashed red line).

To ask whether Epi-MARA's activity prediction, which is based on computationally predicted REST sites, is confirmed by REST binding sites that are indeed occupied by the factor, we mapped REST binding at the ES and NP stages. We carried out ChIP of REST bound DNA and subjected the precipitated DNA to high-throughput sequencing (ChIP-seq). Peak finding was done on pooled replicates and revealed 1599 REST binding peaks in ES cells and 1035 in neuronal progenitors. Identified binding sites show a large overlap to those previously reported (Johnson et al., 2007). The total number of peaks is reduced in progenitors, which likely reflects the fact that REST protein levels decrease during neuronal differentiation (Figure 3.2). In agreement with this hypothesis 97% of the peaks present in progenitors are also present in stem cells. The majority (55%) of REST peaks contain a canonical REST binding site (Table 3.2) in agreement with previous observations.

In addition to the expected enrichment of REST sites at REST binding peaks, there is a significant correlation between the number of predicted sites and the amount of binding as assayed by ChIP-seq (r=0.48, p-value = $2.9 * 10^{-53}$). That is, REST peaks with multiple predicted REST binding sites in general show a stronger ChIP signal. Notably, Epi-MARA takes such binding site multiplicity into account as N_{pm} in equation 3.1 sums up the probabilities of all predicted binding sites at each promoter.

Binding site type	Percentage
Canonical	55
Non-canonical spacing	15
Half-site only	5
No site predicted	25

Table 3.2: REST binding site predictions for REST ChIP-seq data: The majority of REST peaks contain a match to the canonical Rest motif in agreement with previous observations (Johnson et al., 2007). About one quarter of the REST peaks lack a computationally identifiable REST binding site.

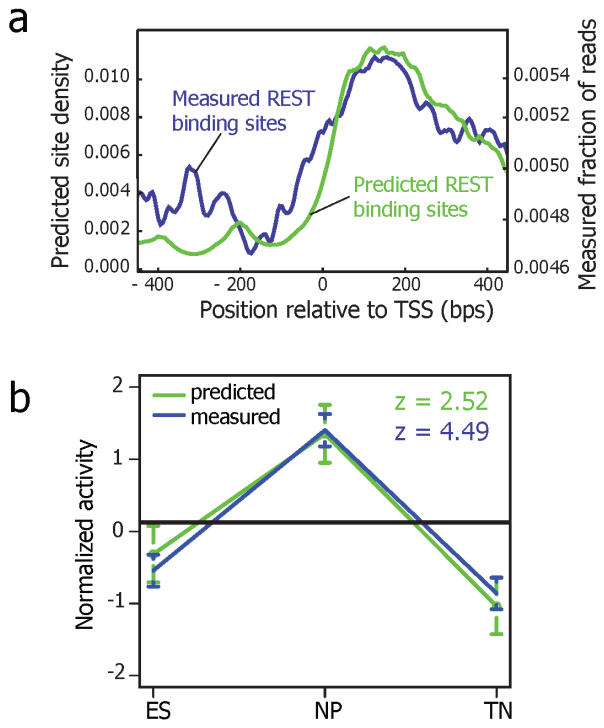

Figure 3.4: **a**) Frequency of predicted (green line) and measured (blue line) binding sites around transcription start sites. **b**) REST activity profiles calculated by Epi- MARA are similar when using either computationally predicted (green line) or measured REST binding sites (blue line). The prediction has higher significance when using the measured sites as indicated by the higher z-value (i.e. higher variance in activity relative to the error-bars).

REST binding peaks are highly enriched in proximity to transcription start sites (TSS) (Figure 3.3) and we classified genes with REST binding within +/- 2 kb of the TSS as potentially regulated by this factor. This resulted in a total of 380 target genes in ES cells and 284 in progenitors, with a 96% overlap. Mammalian promoters are known to separate into two classes associated with either high or low density of CpG dinucleotides (Balwierz et al., 2009; Bird, 1986; Carninci et al., 2006) and we observe that REST predominantly targets high-CpG promoters (Table 3.3). This is consistent with the observation that Polycomb is recruited predominantly to high-CpG promoters (Mendenhall et al., 2010), and that REST-bound regions show increased H3K27me3 levels (Zheng et al., 2009). Interestingly, promoter proximal REST binding sites show a distinct positioning immediately downstream of TSS (Sun et al., 2005; Zhang et al., 2006), which we also observe for both predicted and measured REST binding (Figure 3.4a). While there is general agreement between predicted and measured REST binding, not all predicted promoter sites are occupied and some of the

Promoter class	REST target	Non-target
High-CpG	405 (1.8%)	12886 (56.9%)
Low-CpG	93 (0.4%)	9255 (40.1%)

Table 3.3: Numbers and corresponding percentages of promoters in high- and low-CpG classes that are either bound by REST or not. REST predominantly targets high-CpG promoters, as high-CpG promoters are about 3 times more likely to be REST targets than low-CpG promoters.

promoter proximal REST peaks were missed by the computational predictions. We therefore asked whether Epi-MARA predicts different activities for REST if we replace the computationally predicted REST sites with the actual binding data. This analysis resulted in a strikingly identical activity profile for REST, but with much larger significance as the z-value almost doubled (Figure 3.4b). These results support the REST activity profile inferred using the TFBS predictions. Moreover, it exemplifies how actual *in vivo* binding data can be incorporated and suggests that this increases the accuracy of Epi-MARA's inference.

3.3 REST binding is associated with H3K27me3 dynamics genome-wide

Previous analyses were preformed on H3K27me3 ChIP-chip data, which is limited to assessing H3K27me3 dynamics only at promoters. To analyse H3K27me3 dynamics beyond promoter regions we performed ChIP-seq at the three differentiation stages and determined all genomic regions that were enriched for H3K27me3 in at least one of the cellular states. Although many H3K27me3 enriched regions occur proximal to promoters, more than two thirds of H3K27me3 enriched regions are in fact distal to promoters. However, these distal H3K27me3 regions are much less likely to be targeted by REST than promoter-proximal regions (Table 3.4). Given REST's preferred targeting to high-CpG promoters (Table 3.3), we investigated the CpG content of all H3K27me3 regions genome-wide and found that

H3K27me3 region class	REST target	Non-target
Proximal	351 (1.9%)	5250 (28.7%)
Distal	199 (1.1%)	12496 (68.3%)

Table 3.4: Numbers and corresponding percentages of H3K27me3-enriched regions that are either proximal or distal to a TSS, and that are either targeted by REST or not (Non-target). REST predominantly targets proximal regions with H3K27me3, as proximal regions are 4 times more likely to be a REST target than distal regions.

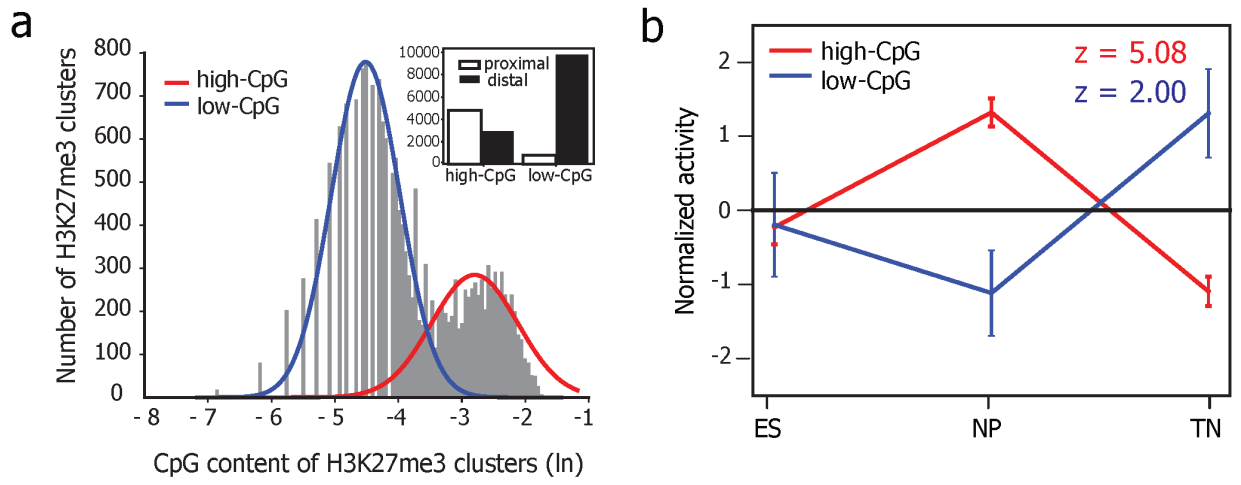

Figure 3.5: **a**) The distribution of CpG frequencies of H3K27me3 regions genome-wide is bimodal and can be fit by a mixture of two log-normal distributions (red and blue lines) corresponding to high- and low-CpG regions, respectively. The inset shows the numbers of H3K27me3 regions that are promoter-proximal and distal for high-CpG and low-CpG regions. **b**) REST activity profiles for high- (red) and low-CpG regions (blue) as inferred by running Epi-MARA on H3K27me3 regions genome-wide show a transient gain and loss, respectively, at the NP stage.

these, like promoters, clearly separate into a high- and low-CpG class (Figure 3.5a). Moreover, CpG content cleanly distinguishes proximal and distal H3K27me3 regions, with 85% of proximal regions being high-CpG and 75% of distal regions being low-CpG (Figure 3.5a). High-CpG regions also show higher levels of H3K27me3 than low-CpG regions (Figure 3.6). Motivated by the above differences between high- and low-CpG regions, and their different sequence composition, we asked whether Epi-MARA predicts different motif activities for REST if we analyse high- and low-CpG regions separately. For high-CpG regions Epi-MARA predicts the same general activity profile for REST as previously for promoters but with even higher significance (Figure 3.5b). For low-CpG regions REST's significance is not only much reduced, but the inferred activity, with a transient loss of H3K27me3 at the NP stage, is almost opposite to that of REST on high-CpG regions (Figure 3.5b). Interestingly, high- and low-CpG regions have distinct H3K27me3 dynamics in general. High-CpG re-

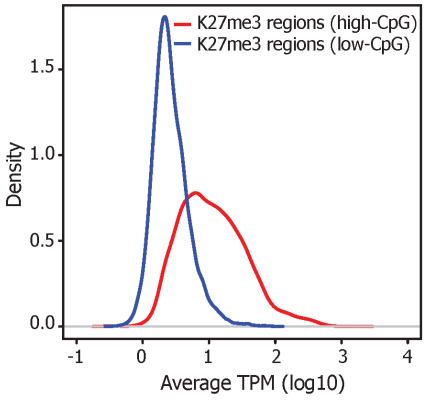

Figure 3.6: Distributions of absolute H3K27me3 levels at all high-CpG (red) and low-CpG regions (blue) that are significantly enriched for H3K27me3.

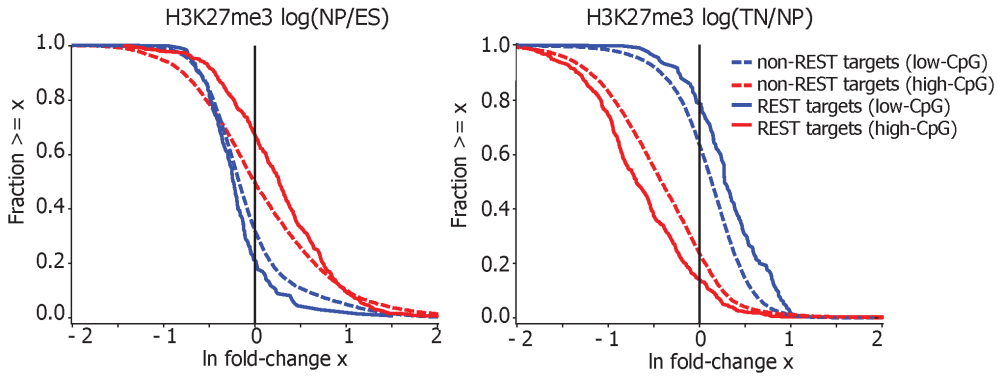

Figure 3.7: Reverse cumulative distributions of changes in H3K27me3 levels at the transition from ES to NP stage (**left**) and NP to TN (**right**). Regions that are enriched for H3K27me3 were divided into high-CpG/low-CpG (red/blue) and REST-target/non-target (solid/broken lines) regions. From the ES to NP transition (left) high-CpG REST targets tend to gain H3K27me3 going from the ES to NP stage whereas non-target high-CpG regions are equally likely to gain or lose H3K27me3. In contrast, low-CpG regions generally lose H3K27me3 and REST targets tend to lose even more H3K27me3. At the transition from the NP to TN stage high-CpG regions generally lose H3K27me3 and REST targets tend to lose even more, whereas low-CpG regions mostly gain H3K27me3 and REST targets gain even more.

gions are more likely to gain H3K27me3 as cells differentiate from the ES to the NP state, than low-CpG regions, which tend to lose H3K27me3 during this transition. The reverse is happening at the transition from the NP to the TN state. Here, high-CpG regions lose H3K27me3,whereas low-CpG regions generally gain H3K27me3 (Figure 3.7). Furthermore, Epi-MARA's predictions for REST are supported by these H3K27me3 dynamics at measured REST sites (Figure 3.7). In summary, genome-wide analysis of H3K27me3 levels predicts that REST binding at high-CpG regions, which includes most promoter proximal regions, leads to a gain in H3K27me3 at the NP stage. In addition, a less significant transient loss of H3K27me3 at the NP stage for low-CpG regions is also predicted by Epi-MARA. We next tested these predictions by analysing cells in which the *Rest* gene had been deleted.

3.4 REST protein is required for local H3K27me3 levels

As already discussed in section 1.2.1 REST is an essential protein for development. However, as knockout ES cells (RESTko) are viable and show no defects in pluripotency (Jorgensen et al., 2009; Yamada et al., 2010) we analysed if they are competent to undergo neuronal differentiation in our *in vitro* system. Here, RESTko cells formed morphologically normal neurons with high efficiency. Moreover, key marker proteins such as Pax6, Nestin and Tuj1

(Heins et al., 2002; Tohyama et al., 1992; Menezes and Luskin, 1994) show similar staining patterns in immunocytochemistry in REST knockout and wildtype (WT) cells (Figure 3.8a).

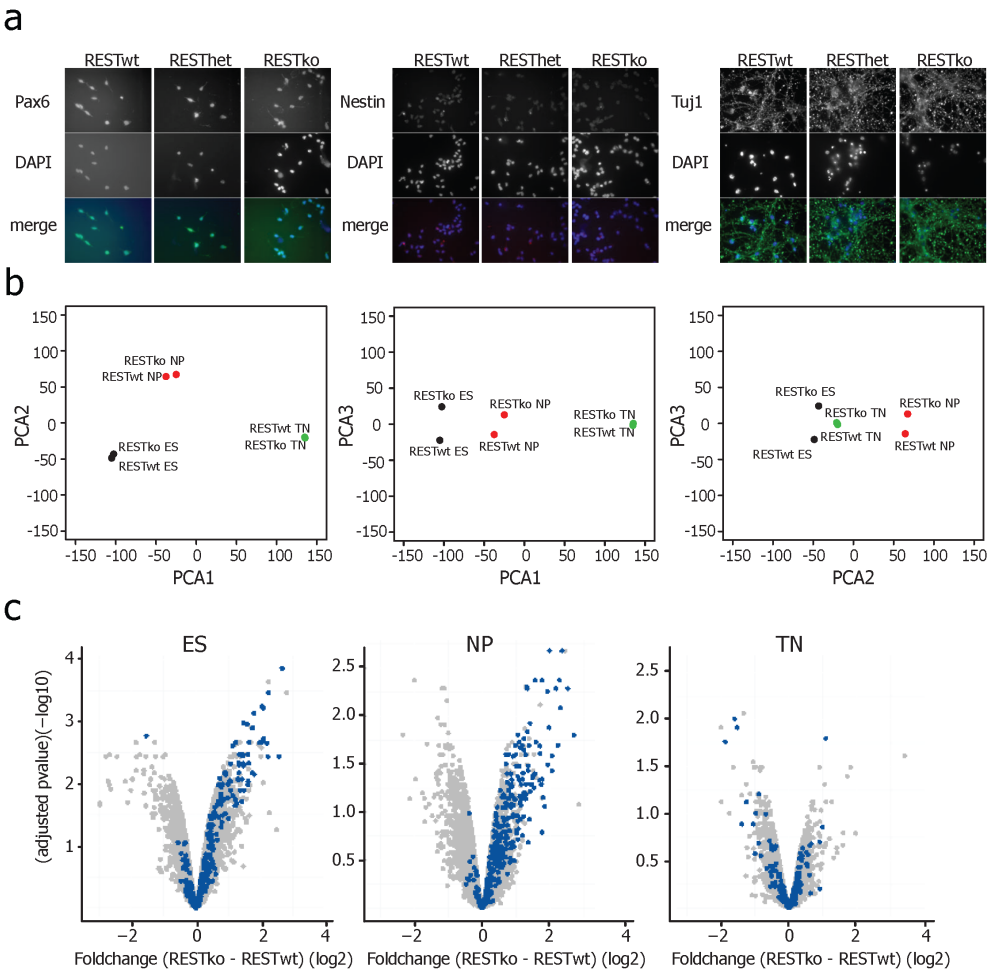

Figure 3.8: **a**) Marker proteins show similar staining patterns in immuno-cytochemistry in REST knockout and wildtype (WT) cells: REST wildtype (RESTwt), heterozygous (RESThet) and homozygous knockout (RESTko) neuronal progenitors and terminal neurons were fixed and stained for several marker proteins specific for the NP stage (Pax6 (top panel) and Nestin (middle panel)) (Heins et al., 2002; Tohyama et al., 1992) and TN stage (Tuj1 (bottom panel))(Menezes and Luskin, 1994), respectively. The cells shown are representative for the population. **b**) Principal component analysis of the gene expression profiles of wildtype (RESTwt) and REST knock out (RESTko) cells. RESTko cells cluster with the corresponding wildtype stage. **c**) Volcano plots depict, for each gene, the fold-change in gene expression in RESTko vs RESTwt cells and the corresponding adjusted p-value for all three stages of differentiation. REST target genes are colored in blue.

Class and Stage	Percentage that lose H3K27me3	Percentage that gain H3K27me3
ES (low-CpG)	4.9% +/- 3.2%	2.9% +/- 2.3%
ES (high-CpG)	5.2% +/- 2.3%	0.9% +/- 0.8%
NP (low-CpG)	1.4% +/- 4.4%	12.4% +/- 4.4%
NP (high-CpG)	21.7% +/- 2.8%	0.8% +/- 0.7%

Table 3.5: Estimated percentages of REST targets that significantly lose/gain H3K27me3 in the RESTko cells, separately for low- and high-CpG regions, and for the ES and NP stages. Error bars are based on a Bayesian inference procedure. The strongest changes are observed for high-CpG regions losing H3K27me3 at the NP stage, followed by low-CpG regions gaining H3K27me3 at the same stage.

We also carried out transcriptome analysis of REST wildtype and REST knockout cells during *in vitro* neuronal differentiation, which confirmed that only a relatively small number of genes show significant expression differences in the absence of REST (Figure 3.8b,c). Importantly, many of these are direct targets of REST that are up-regulated in the knockout (Figure 3.8c). We conclude that REST is not essential for the initial steps of neuronal differentiation *in vitro* and thus measured genome-wide H3K27me3 levels in RESTko cells at the stem cell and progenitor stages to investigate whether the absence of REST affects H3K27me3 levels at its target genes. For subsequent analysis, we separated all regions enriched for H3K27me3

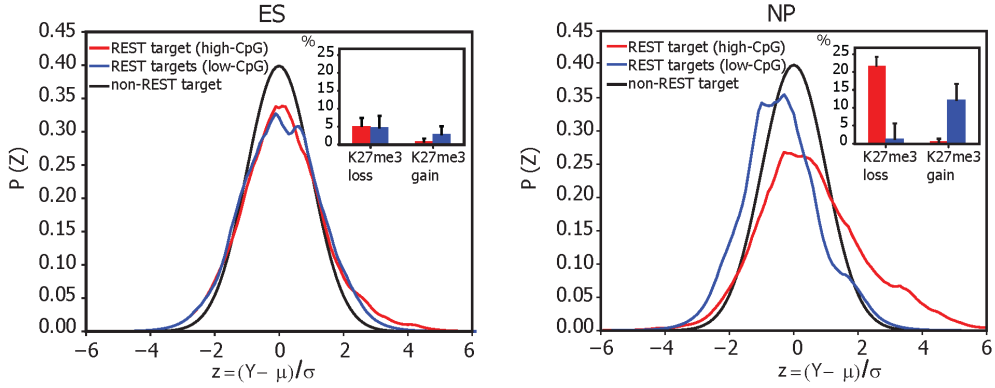

Figure 3.9: Global comparison of H3K27me3 levels between WT and RESTko cells. Shown are the normalised distribution of the ratio between H3K27me3 in WT versus RESTko for non-target regions (black lines) and for either low-CpG (blue lines) or high-CpG (red lines) regions that are REST targets at the ES (**left panel**) and NP (**right panel**) stage. The insets show the estimated fractions of REST targets that significantly lose or gain H3K27me3 in the RESTko at high-CpG (red) and low-CpG regions (blue).

at any of the stages into high-CpG and low-CpG and further into REST-target and non-targets. Next, we compared H3K27me3 levels in wildtype and RESTko cells between these four classes. This reveals little difference between REST target regions and non-target regions at the ES stage (Table 3.5 and Figure 3.9), in line with Epi-MARA's predicted REST activity

at this stage. In contrast at the NP stage, as exemplified at two loci in Figure 3.10, we observe a substantial loss of H3K27me3 in the RESTko cells relative to wildtype cells, affecting a substantial number of high-CpG REST targets (Table 3.5 and 3.9). In addition, although the changes at low-CpG regions are much weaker, a notable gain of H3K27me3 is observed at low-CpG REST targets (Figure 3.9). This experimentally confirms Epi-MARA's predictions for REST at both high- and low-CpG regions. We conclude that REST contributes functionally to local levels of H3K27me3, in particular at high-CpG regions in neuronal progenitors.

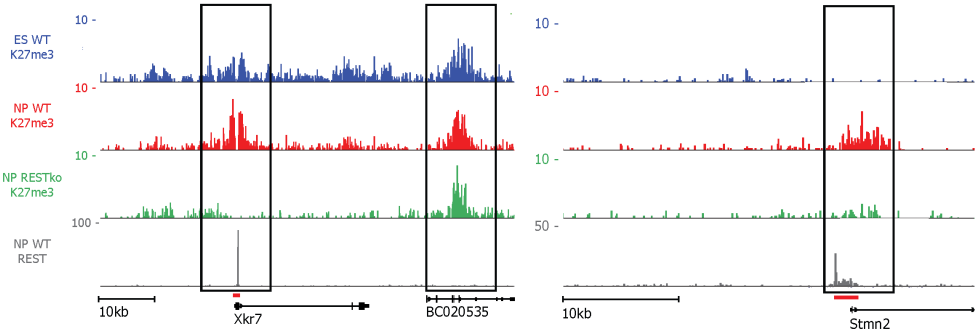

Figure 3.10: ChIP-Seq signal for H3K27me3 and REST in representative genomic regions. Shown are H3K27me3 signals in ES cells, NPs of wildtype (WT) and RESTko cells as well as REST signal in NPs. The left panel exemplifies selective loss of H3K27me3 at the REST binding site of the Xkr7 locus, whereas neighbouring regions (BC020535) remain unaffected. The right panel shows similar loss of H3K27me3 at the Stmn2 locus. Both the Xkr7 and Stmn2 locus are examples of promoter proximal high-CpG regions. Shown are normalised read densities. The red bars at the REST peaks indicate the regions cloned for transgenic experiments.

We further tested, if the observed loss of H3K27me3 is accompanied by a loss of PRC2, which mediates the H3K27me3 mark. To this end we compared occupancy of the PRC2 component SUZ12 in RESTwt and RESTko neuronal progenitors.

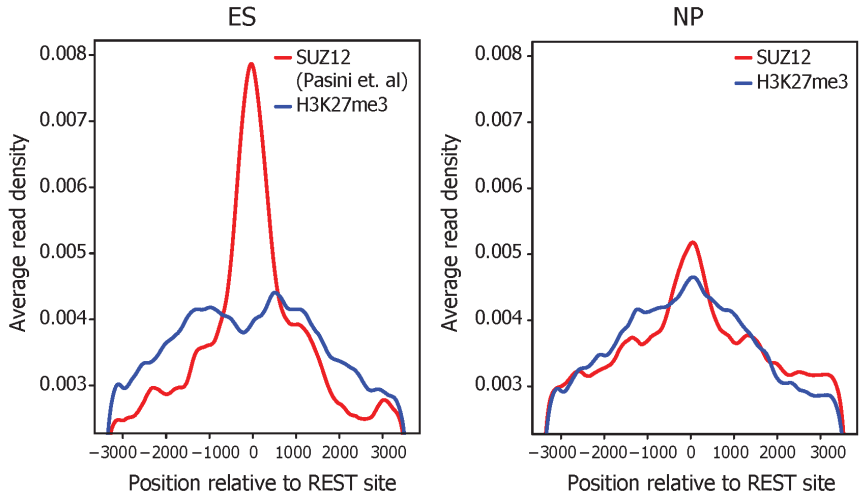

Figure 3.11: H3K27me3 and SUZ12 levels peak around REST binding sites. Shown is the normalised average read density of SUZ12 (red) and H3K27me3 (blue) in ES cells (Pasini et al., 2010) and neuronal progenitors.

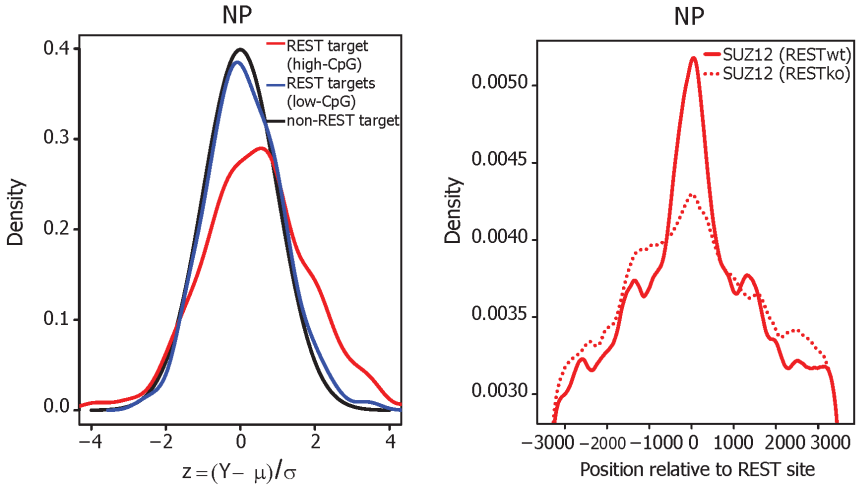

Figure 3.12: **a**) Comparison of SUZ12 levels between WT and RESTko cells in H3K27me3-enriched regions at the NP stage. Shown are the distributions of the normalised difference in SUZ12 levels (represented as a z-statistic) in WT versus RESTko for non-target regions (black line) and for REST targets in either low-CpG (blue line) or high-CpG (red line) regions. Few low-CpG REST targets significantly change, whereas a considerable fraction of high-CpG REST targets show loss of SUZ12 in the RESTko cells. **b**) Absence of REST reduces the localisation of SUZ12 at REST peaks. Shown is the normalised average read-density of SUZ12 in WT and RESTko neuronal progenitors.

Interestingly, both H3K27me3 and SUZ12 are enriched at REST binding sites in both ES and NPs (Figure 3.11). More importantly we observe a loss of SUZ12 at a substantial number of high-CpG REST targets (Figure 3.12a) as well as a loss of co-localisation of

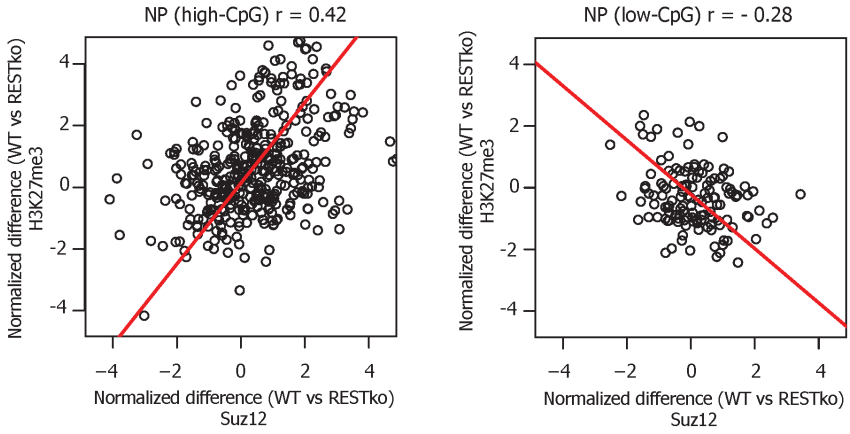

Figure 3.13: Comparison of the effects on SUZ12 and H3K27me3 levels of REST knockout at REST target regions. For each REST target the normalised difference (z-statistic) between WT and RESTko levels for SUZ12 (x-axis) and H3K27me3 (y-axis) are shown at the NP stage for high-CpG regions (**left**) and low-CpG regions (**right**). Significant correlation is observed between loss of H3K27me3 and loss of SUZ12 cells for high-CpG REST targets in RESTko cells (r=0.42, p-value $< 2.23 * 10^{-16}$). A weak anti-correlation is observed for low-CpG REST targets (r=-0.28, p-value < 0.001)

SUZ12 with REST binding (Figure 3.12b). Moreover, compatible with a role for REST in Polycomb recruitment, there is a correlation between reduction in SUZ12 levels and reduction in H3K27me3 levels at high-CpG REST targets (Figure 3.13 left). A weak anti-correlation is observed for low-CpG REST targets (Figure 3.13 right).

3.5 REST affects H3K27me3 and gene expression independently

REST is an established repressor of gene activity and a subset of REST target genes are transcriptionally up-regulated in RESTko cells (Figure 3.8c). It is thus conceivable that the observed loss of H3K27me3 at proximal REST targets is a direct consequence of transcriptional up-regulation. This in turn would imply that all genes with REST-dependent loss of H3K27me3 are transcriptionally up-regulated in RESTko cells. However, a direct comparison reveals only a weak correlation (r=0.28 in ES and r=0.44 in NP) between changes in H3K27me3 at proximal regions and changes in gene expression (Figure 3.14). At the ES stage many REST targets are transcriptionally up-regulated without showing a loss of H3K27me3. Most importantly, a third of the regions that lose H3K27me3 at the NP stage are not significantly up-regulated in expression (Figure 3.14a). We thus conclude that the crosstalk between REST and the Polycomb pathway is independent of transcriptional changes at a substantial number of REST targets. This does not imply that there is no association at

all between changes in H3K27me3 levels and expression changes. Indeed, as expected given that H3K27me3 is a known repressive mark, we observe that targets that lose H3K27me3 in the RESTko are more likely to be up-regulated transcriptionally than those that do not lose H3K27me3 (Figure 3.14b).

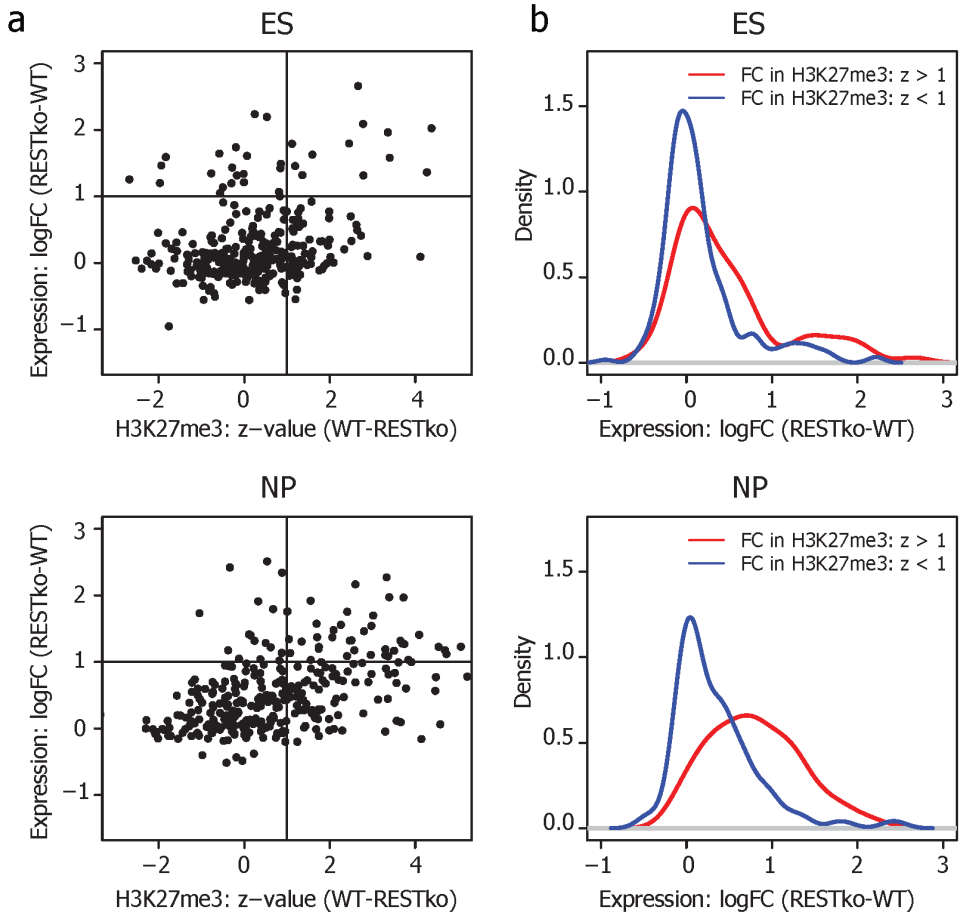

Figure 3.14: **a**) Pairwise comparison of changes in H3K27me3 levels (z-value wildtype - RESTko) versus changes in transcription (log fold-change RESTko - wildtype) at the ES (top panel) and NP stage (bottom panel). The horizontal and vertical lines correspond to a z-value of 1 and a log fold-change of 1 (i.e. two-fold up-regulation). **b**) Distribution of expression log fold-changes (RESTko - wildtype) for REST targets that significantly lose H3K27me3 (z>1, red lines) and REST targets that do not significantly lose H3K27me3 (z<1, blue lines), both at the ES (top panel) and NP (bottom panel) stage.

3.6 Promoter fragments containing TFBS recruit H3K27me3

Having established that the absence of REST protein leads to a decrease of H3K27me3 at a significant fraction of its high-CpG binding sites, we wanted to further ask whether fragments of high-CpG promoter regions containing a REST binding site can autonomously recruit H3K27me3 and whether the REST site contributes to this recruitment.

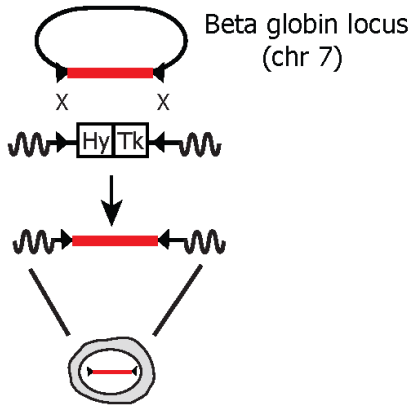

Figure 3.15: Strategy to insert promoter regions into a defined genetic site (beta globin locus) via recombinase-mediated cassette exchange. The two inserted marker genes confer resistance against hygromycin (Hy) and sensitivity against ganciclovir (Tk), respectively and are flanked by two inverted lox sites (black triangles). Targeted insertion of a given transgene is achieved by Cre-mediated recombination and negative selection.

To this end we generated reporter constructs consisting of 1.2 to 2.5 kb promoter fragments containing a REST site, and mutant versions in which the REST site had been deleted. To ensure proper chromatin organisation we placed these sequence variants in wildtype cells into the same chromosomal locus using recombinase-mediated cassette exchange (RMCE) as depicted in Figure 3.15 (Lienert et al., 2011b; Feng et al., 1999). This site-specific targeting controls for the genomic environment and thus allows direct comparison of wildtype and mutant sequences. Importantly, the chosen "test site" is positioned within a genomic region that harbours no H3K27me3 and no REST binding sites (Lienert et al., 2011b; Stadler et al., 2011). Thus, any REST or H3K27me3 signal should primarily reflect the recruitment abilities of the inserted sequence. We inserted wildtype and mutated (ΔREST) promoter fragments (Figure 3.16 a) of the following genes: *Stmn2*, *Xkr7*, *Bdnf* and *Pgbd5* (the size of the cloned promoter fragments for *Stmn2* and *Xkr7* is depicted by a red bar in Figure 3.10). After targeted insertion and differentiation into neuronal progenitors we detect strong REST binding by ChIP to the wildtype, but no or weak binding to the four ΔREST mutant sequences showing that the REST site is essential for REST binding to the reporter constructs (Figure 3.18).

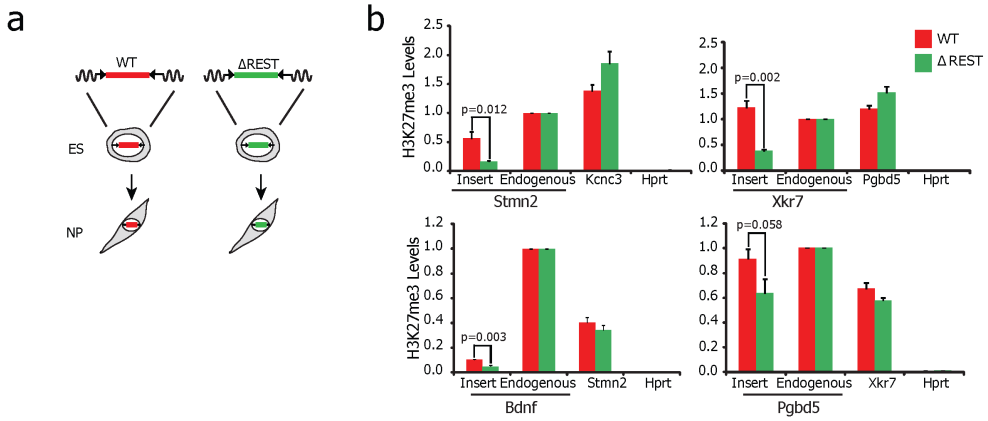

Figure 3.16: **a**) The RMCE approach was used to insert several REST target promoter fragments with either wildtype sequence (WT) or REST site mutation (ΔREST) into the beta globin locus. Correctly targeted ES cells were differentiated to the NP stage, where H3K27me3 and REST were measured at the inserted fragments. **b**) For each of the 4 inserts H3K27me3 levels were measured in cells bearing the WT fragment (red bars) and in cells bearing the ΔREST fragment (green bars). Levels were measured at, from left to right in each panel, the inserted region, the corresponding endogenous locus, a positive control, and a negative control region. All H3K27me3 levels are scaled to that of the endogenous region and error-bars show the standard error of three biological replicates. A p-value is shown and calculated for each insert using unpaired one-tailed t-test statistics.

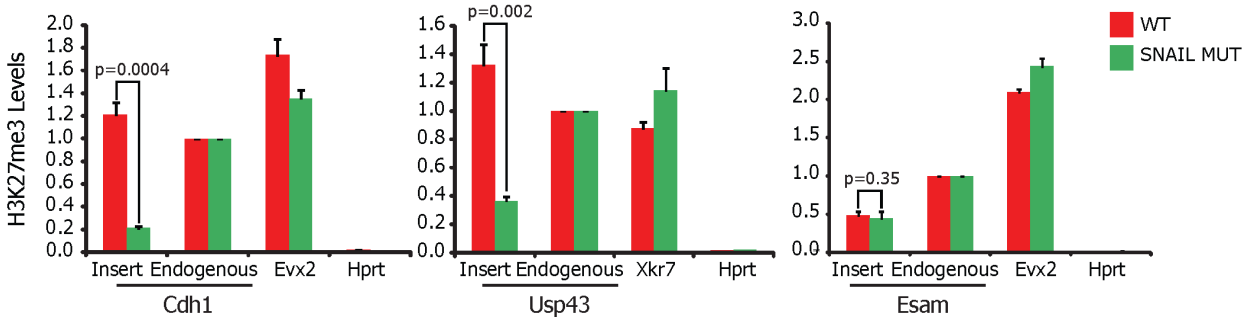

Figure 3.17: **a**) Either wildtype (WT) or mutated (MUT) promoter regions containing predicted SNAIL sites were inserted via RMCE. The SNAIL sites were mutated by changing the first and last nucleotide of the motif to a Thymidine. Correctly targeted ES cells were differentiated to the NP stage. **b**) For each of the 3 inserts H3K27me3 levels were measured in cells bearing the WT (red bars) or mutated promoters (green bars), respectively. Note that the *Cdh1*, *Usp43* and *Esam* promoter regions have three, two and one predicted/mutated SNAIL site, respectively. Levels were measured at, from left to right in each panel, the inserted region, the corresponding endogenous locus, a positive control, and a negative control region. All H3K27me3 levels are scaled to that of the endogenous region and error-bars show the standard error of three biological replicates. A p-value is shown and calculated for each insert using unpaired one-tailed t-test statistics.

Importantly, H3K27me3 is observed at all promoter fragments containing a functional REST site at the progenitor stage, whereas the mutant sequences show significant loss of H3K27me3 (Figure 3.16 b). In case of the Stathmin-like 2 (Stmn2) promoter presence of the REST site results in a more than three-fold increase of H3K27me3 signal. Notably, the endogenous Stmn2 promoter shows no transcriptional response in RESTko cells. Of all four tested promoter fragments the Pgbd5 fragment shows the weakest loss of H3K27me3. Importantly however, the corresponding loss of REST binding at this promoter is also the weakest (Figure 3.18), suggesting that a cryptic binding site may still remain at this fragment.

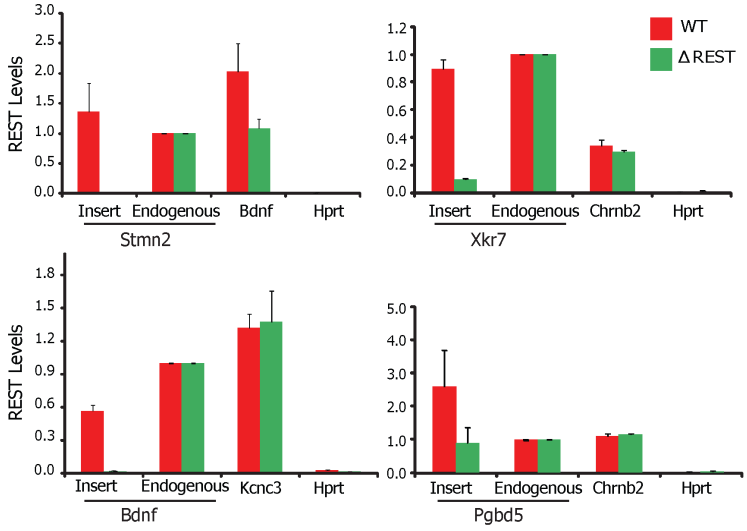

Figure 3.18: Transgenic wildtype promoters show strong REST binding, but no or weak binding at the four ΔREST sequences. Levels were measured at, from left to right in each panel, the inserted region, the corresponding endogenous locus, a positive control, and a negative control region. All REST levels are scaled to that of the endogenous region and error-bars show the standard error of three biological replicates.

In summary, we conclude that promoter fragments containing a REST binding site are sufficient to recruit H3K27me3 and that the REST binding site itself is a major contributor *in cis* to these H3K27me3 levels. Together with the observed changes in H3K27me3 levels at genome-wide REST targets in the RESTko cells this firmly establishes that REST binding mediates Polycomb targeting and contributes to local levels of H3K27 methylation.

Besides REST, many factors that Epi-MARA predicts to play a role in H3K27me3 dynamics are recognised by a family of TFs. This makes loss of function approaches at the protein level very demanding. Our transgenic approach, however, can be used to assess the contribution of binding motifs to Polycomb recruitment irrespective of which TF from a family is binding. We thus extended our analysis to study the effect of the SNAIL binding site, another motif predicted to recruit K27me3 at the NP stage (Figure 3.1 left). We inserted a total of six regulatory regions containing wildtype or mutated Snail motif (Figure 3.17) and tested for presence of H3H3K27 methylation. As seen with regulatory regions containing Rest motifs, we observe that all constructs containing Snail motifs are sufficient to recruit H3K27me3. Deletion of Snail motifs leads to significant reduction of H3K27me3 for two of the three constructs tested

(Figure 3.17). Notably, the construct that showed no significant response was the only one that contained only a single predicted SNAIL site, suggesting that the effect on H3K27me3 increases with the number of sites.
In summary, we conclude that promoter fragments containing bindings sites for SNAIL and REST TFs are sufficient to recruit H3K27me3 and, in line with the predictions, that these binding sites are a major contributor *in cis* to H3K27me3 levels.

3.7 RESTko NPs show increase in H3K4 methylation

In the last sections we confirmed experimentally that REST binding to promoter-proximal high-CpG regions is required for an increase in H3K27 methylation as ES cells differentiate to neuronal progenitors. This process is to a significant extent independent of transcriptional up-regulation (see section 3.5) and more importantly depends on local DNA sequence (see section 3.6) further supported by the results obtained from the mutation of predicted Snail motifs (Figure 3.17). As presented in section 1.3.3 many mechanisms have been proposed to locally

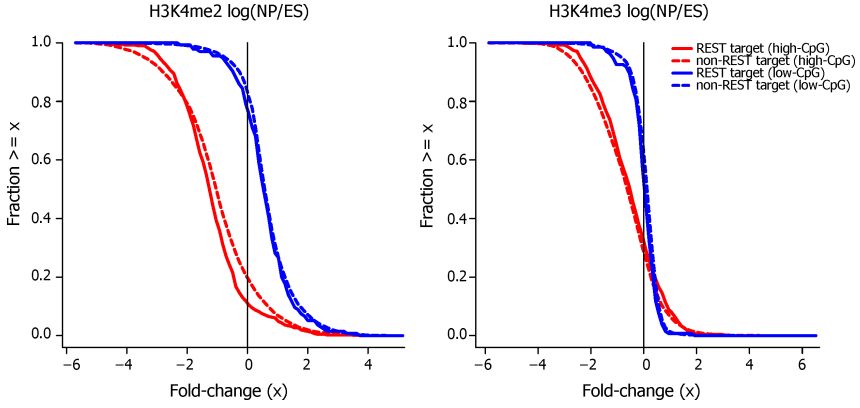

Figure 3.19: Reverse cumulative distributions of changes in H3K4me2 (**left**) and H3K4me3 ((**right**) levels at the transition from ES to NP stage. Regions were classified as before according to their H3K27me3 enrichment and divided into high-CpG/low-CpG (red/blue) and REST-target/non-target (solid/broken lines) regions. High-CpG REST- and non-REST targets show a strong loss of both H3K4me2 and H3K4me3 at the transition from ES to NP.

recruit Polycomb. The fact that we predict multiple TFs to regulate a gain in H3K27me3 levels from the transition of ES cells to NPs does not favour a model of direct interaction between Polycomb and REST or SNAIL proteins (see section 5.2 of the discussion). Of the motifs inferred by Epi-MARA that explain most of the H3K27me3 changes during neuronal differentiation (Figure 3.1 left) many show a transient gain of H3K27me3 at the NP stage,

which is in agreement with the fact that about half of all high-CpG H3K27me3 regions gain H3K27me3 from the ES to the NP stage (Figure 3.7). Interestingly, this gain of H3K27me3 at high-CpG regions is accompanied by a strong decrease in H3K4me2 and H3K4me3 (Figure 3.19).

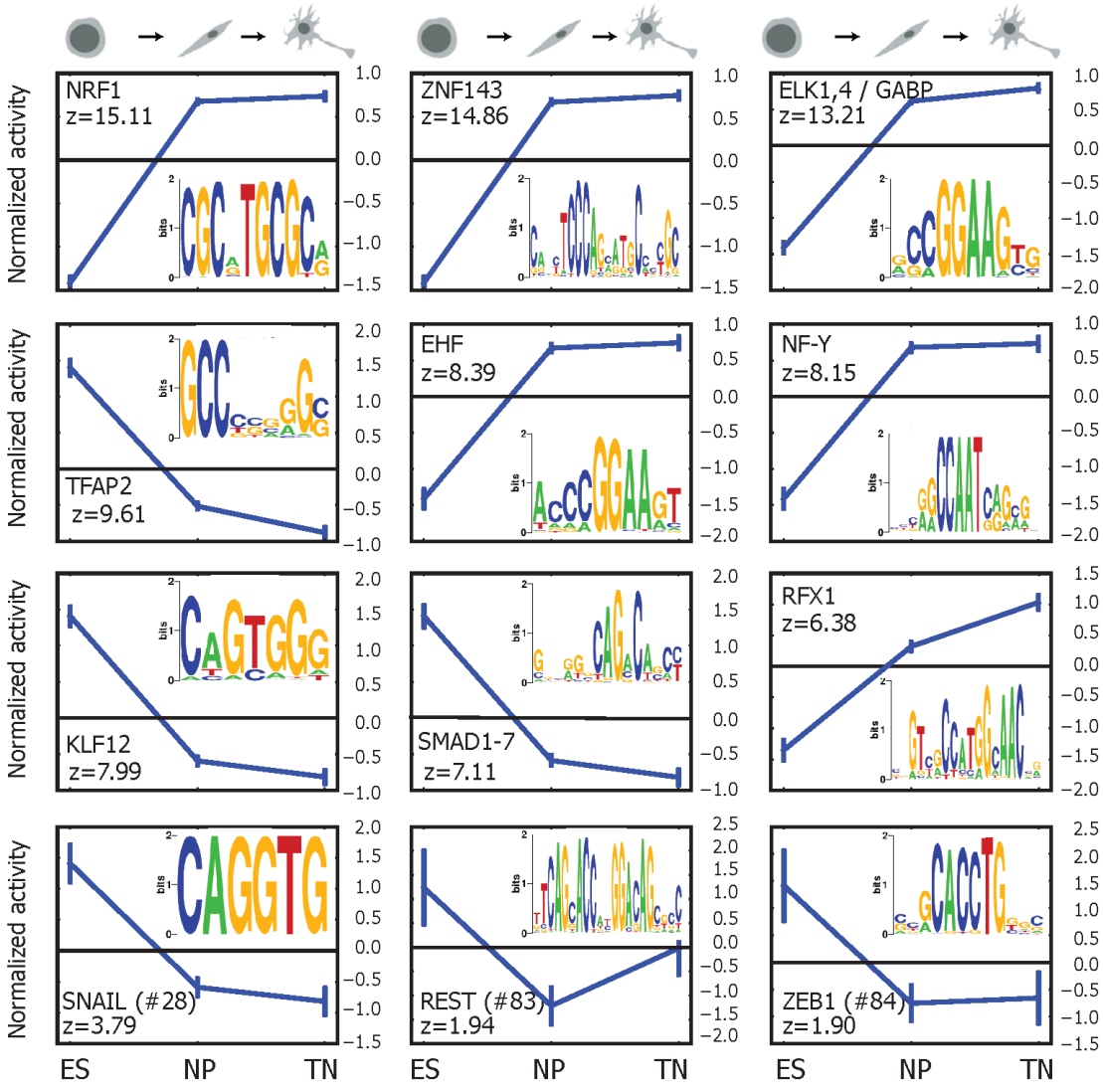

Figure 3.20: Epi-MARA predicts transcription factors that explain dynamics in H3K4me2 levels during neuronal differentiation. Depicted are the normalised activity profiles of the top nine motifs (blue lines, with standard errors indicated) with their respective z-values as well as Epi-MARA predictions for the TFs REST, SNAIL and ZEB1. The time points correspond to the embryonic stem cell (ES), neuronal progenitor (NP), and terminal neuron (TN) stage. Sequence logos of each of the motifs and the transcription factors thought to bind to them are shown as insets.

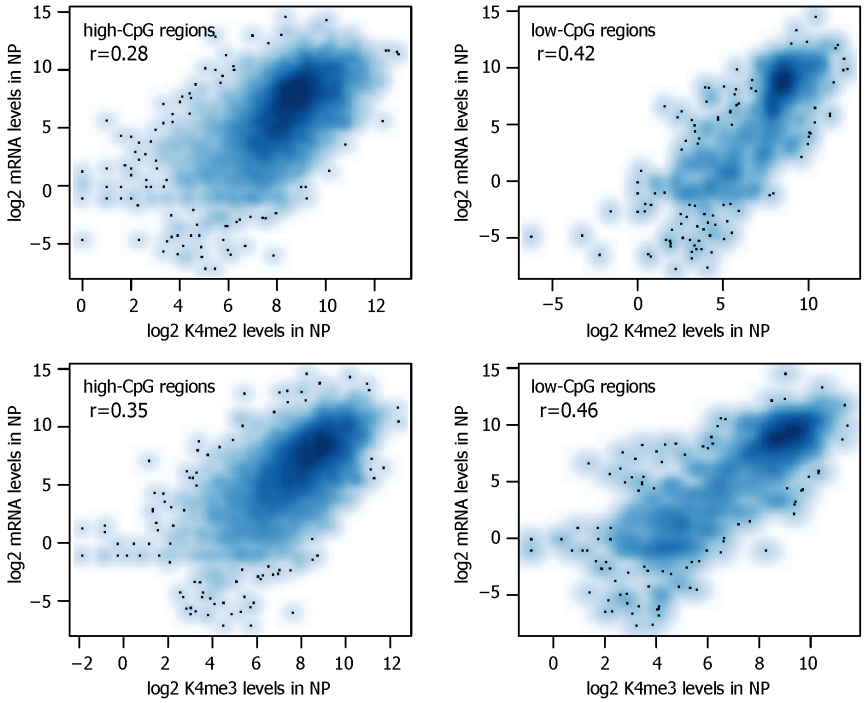

Figure 3.21: Scatter plots comparing levels of H3K4me2 (**top**) and H3K4me3 ((**bottom**) (measured at +-2kb around TSS) to levels of mRNA (measure over the entire gene). The selected regions correspond to high-CpG ((**left**) and low-CpG ((**right**) H3K27me3 regions that overlap promoters (+-2kb around TSS). The correlation coefficient was calculated by Pearson correlation and is generally higher for low-CpG regions than for high-CpG regions.

As mentioned in section 1.3.5 a recent study indicated that methylation of H3K4 inhibits the activity of PRC2 to methylate H3K27 (Schmitges et al., 2011), suggesting that the observed changes in H3K27me3 and H3K4me2/3 might be functionally connected. Interestingly, the genome of *Drosophila melanogaster*, which is mostly devoid of DNA methylation and CpG Islands, shows a mutually exclusive pattern of H3K27me3 and H3K4me3 (Tolhuis et al., 2006). Furthermore, REST was shown to interact with lysine-specific demethylase 1 (LSD1) (Lee et al., 2005) suggesting that REST might reduce local H3K4me2/3 levels that allow subsequent methylation of H3K27. Supporting a role for REST in H3K4 demethlyation are Epi-MARA predictions for TFs that explain dynamics in H3K4me2 during *in vitro* neuronal differentiation (Figure 3.20). Rest is not among the top motifs explaining changes in H3K4me2 methylation but still shows a significant z-value. Interestingly, Epi-MARA predicts similar activities for Snail, Rest and Zeb-1 motifs for H3K4me2 dynamics with a decrease in motif activity specifically at the NP stage, while at the same time these three motifs are predicted to regulate a gain of H3K27me3 (Figure 3.1 left). The top motifs that explain changes

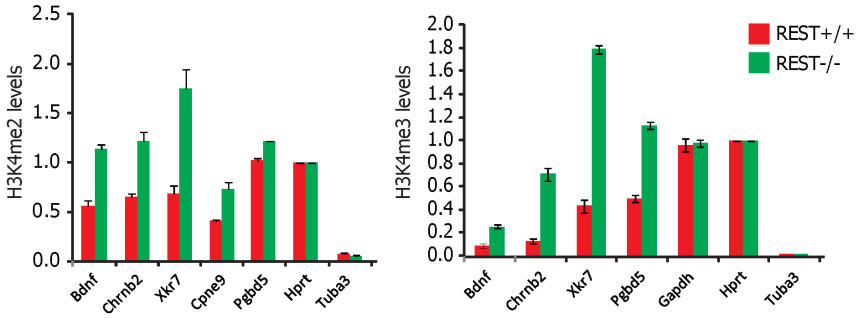

Figure 3.22: Quantitative PCR of H3K4me2 and H3K4me3 ChIPs in REST wildtype and RESTko neuronal progenitors. Enrichments are normalized to a positive control(Hprt). RESTko cells show an increase for both H3K4me2 and H3K4me3 at REST targets (Bdnf, Chrnb2, Xkr7, Pgbd5, Cpne9). GapDH and Hprt serve as positive controls, Tuba3 as a negative control, respectively. Shown are mean enrichments. Error bars show the standard deviation of three biological replicates

in H3K4me2 during neuronal differentiations contain many transcriptional activators such as NRF1, ZNF143, ELK1 and NFY. Interestingly, NRF1 was reported to regulate neurite outgrowth (Chang et al., 2005), which is consistent with an increase in motif activity during neuronal differentiation. As mentioned in the introduction (section 1.3.2) both H3K4me2 and H3K4me3 correlate with increased gene expression. Thus, changes in H3K4 methylation in the RESTko cells might simply be the result of transcriptional up-regulation of REST target genes (see Figure 3.8). Figure 3.21 shows the correlation of H3K4me2 and H3K4me3 to gene expression levels measured by mRNA sequencing in neuronal progenitors. As this project focuses on H3K27me3 enriched regions the data is shown for high- and low-CpG H3K27me3 regions that overlap with promoters only. As mentioned in the introduction (section 1.3.2) H3K4me2 and H3K4me3 levels are more predictive of gene expression at low-CpG regions, which can be seen by the higher correlation coefficient for low-CpG region for both H3K4me2 and H3K4me3 (Figure 3.21).

To further analyse the potential interplay between H3K4 methylation and H3K27me3 in the absence of REST we measured H3K4me2 and H3K4me3 levels in RESTko neuronal progenitors and compared them to wild type levels. We detect an increase of both H3K4me2 and H3K4me3 at several REST binding sites by qPCR (Figure 3.22) and further analyse the differences in H3K4me3 methylation in the absence of REST genome-wide by ChIP-seq. As shown in Figure 3.21 changes in H3K4me3 correlate well with gene expression. Thus, we analysed whether the observed changes in H3K4me3 methylation in RESTko neuronal progenitors are better explained by a change in gene expression or by a change in H3K27me3 (Figure 3.23). The measured increase in H3K4me3 in RESTko cells shows a high positive correlation with

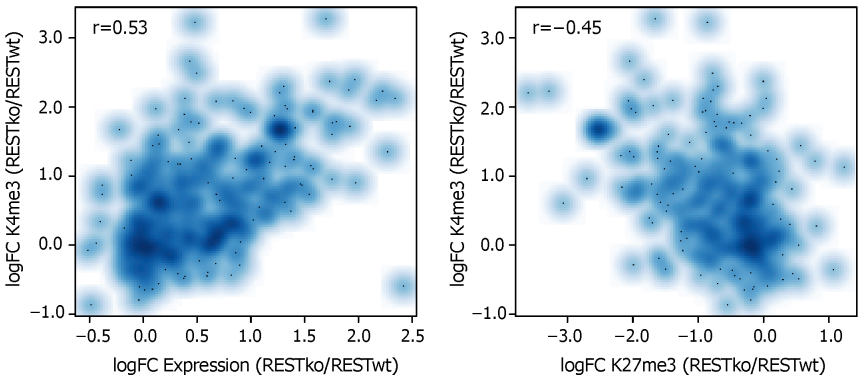

Figure 3.23: Scatter plots comparing the fold changes in H3K4me3 levels in RESTko versus RESTwt to either fold changes observed in gene expression (left) or H3K27me3 levels (right) also in RESTko versus RESTwt. The selected regions correspond to high-CpG REST targets regions that overlap promoters (+-2kb around TSS). As expected H3K4me3 shows a high positive correlation to gene expression and a slightly lower negative correlation to H3K27me3.

transcription changes and a slightly lower negative correlation with the observed H3K27me3 loss in RESTko cells. This suggests that the three signals are tightly linked and makes it difficult to separate them (see discussion 5.2). To further assess if transcript or H3K27me3 levels are a better predictor of H3K4me3 levels in RESTko - RESTwt ($\Delta(KO - WT)$) we built a simple linear model that tries to answer this very question (Equation 3.2):

$$H3K4me3_{(\Delta(KO-WT))} = \alpha RNA_{(\Delta(KO-WT))} + \beta H3K27me3_{(\Delta(KO-WT))} \tag{3.2}$$

Overall, this linear model is able to explain 33% of the observed variance in H3K4me3 between RESTko and RESTwt cells. It assigns gene expression changes ($RNA_{(\Delta(KO-WT))}$) a coefficient of $\alpha = 0.53$ (P-value $< 2*10^{-16}$) and H3K27me3 changes ($H3K27me3_{(\Delta(KO-WT))}$) a coefficient of β = - 0.26 (P-value = $4.47 * 10^{-09}$).
We conclude that gene expression changes are a better predictor of H3K4me3 changes than H3K27me3. However, not everything is explained by transcription as the contribution of H3K27me3 to the linear model is still highly significant. It improves the linear model by explaining differences that were not explained by changes in expression. Thus, further experiments will be required to uncover the causal relationship between transcription, H3K4 and H3K27 methylation.

3.8 REST binding is determined by REST site quality

REST protein levels decrease during neuronal differentiation (Figure 3.2) and only a subset of REST peaks present in ES cells are still occupied in NPs (Figure 3.3 and section 3.2). Thus, besides a role for REST in Polycomb recruitment its binding data can further give insights into how dynamics binding of this TF is regulated during *in vitro* neuronal differentiation. One limitation for such analysis is the fact that even combining all three biological ChIP-seq replicates of ES or NP does not generate saturated REST ChIP-seq data. That is, with an increasing depth of sequencing more and more peaks are called. Thus, the current data do not allow clear conclusions regarding presence or absence calls of ES-specific REST peaks in NPs. Nevertheless, the peaks that are classified as ES-specific are clearly enriched below the cut-off in NPs. Thus, the dynamic binding of REST can still be informative regarding the relative levels of REST at shared peaks between ES and NP and ES-specific peaks. One straightforward explanation for the differences in REST binding would be that shared

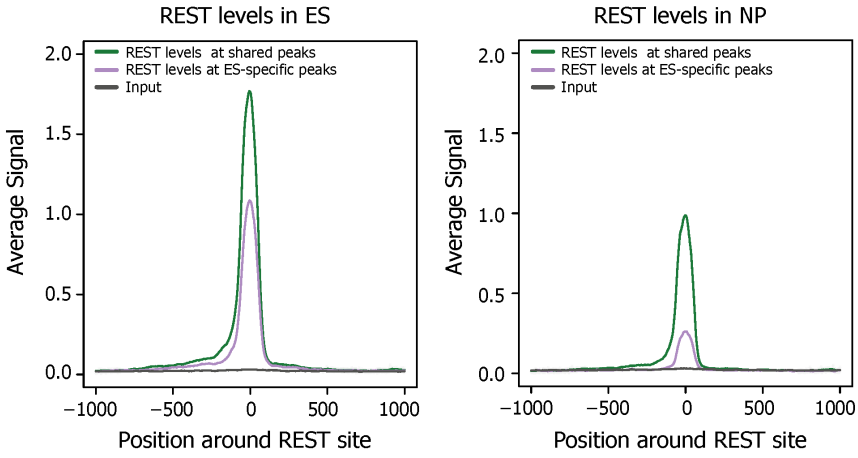

Figure 3.24: Average profile of REST levels at ES-specific and shared REST peaks between ES and NPs. REST levels at ES-specific peaks are already reduced compared to shared peaks in ES cells (left). In neuronal progenitors (right) ES-specific peaks show strongly reduced REST levels close to background.

peaks between ES and NP have a high-affinity REST site that is still bound in neuronal progenitors, whereas ES-specific REST peaks would have REST sites with a lower binding affinity. Supporting such a notion is the fact that ES-specific REST peaks, that are mostly lost in NPs, show already lower REST levels in ES cells (Figure 3.24). Overall REST levels are decreased from the transition from ES cells to neuronal progenitors, which is consistent with the fact that REST protein levels also decrease (Figure 3.2).

To test, if the quality of the REST sites differs between shared and ES-specific REST peaks,

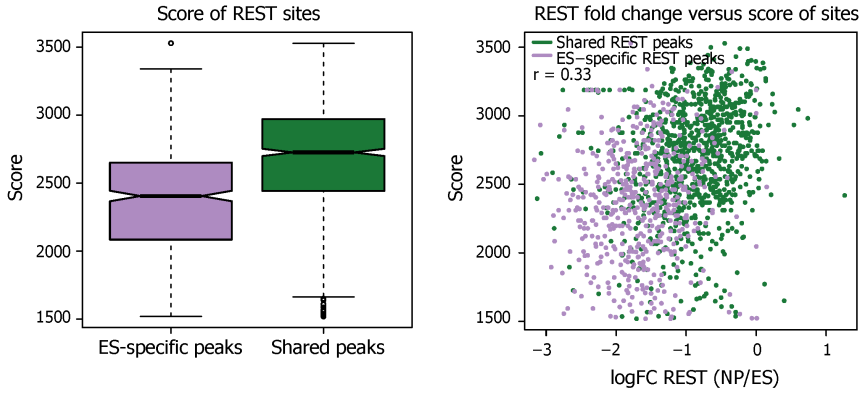

Figure 3.25: Comparison of REST binding site quality of ES-specific and shared REST peaks. **left**) Boxplot shows that shared REST peaks harbour REST sites with a significantly higher score than ES-specific peaks (P-value $< 2.2 * 10^{-16}$). **right**) Scatter plot shows the fold change of REST levels in NP-ES for ES-specific (purple) and shared (green) REST peaks versus the score of the peaks. Peaks with the smallest decrease from ES to NP, tend to have the highest site score (r=0.33).

we calculated a quality score for each REST site. This score reflects how similar the site is to the original position-weight-matrix (PWM), which represents the best possible REST site and thus the highest possible score. Comparing the REST site scores for ES-specific REST peaks versus those that are shared between ES and NP indeed shows that the shared peaks have a significantly higher score (P-value $< 2.2 * 10^{-16}$) compared to the ES-specific peaks (Figure 3.25 left). Even more, there is a significant correlation between the fold change of REST levels between ES and NP and the site score, showing that high affinity sites retain most REST in neuronal progenitors (r=0.33) (Figure 3.25 right). These results suggest that, at least for the TF REST, binding of transcription factors is partly regulated by the quality of binding sites.

Chapter 4

Methods

4.1 Epi-MARA

We describe the main methods employed in the Epi-MARA analysis. Further details are supplied in (Arnold, 2011). Epi-MARA models the dynamics of epigenetic marks in terms of predicted TFBSs in regulatory regions genome-wide, building on the motif activity response analysis that was developed previously by the van Nimwegen Lab (Suzuki et al., 2009). Briefly, using the MotEvo suite (Arnold et al., 2012), for each promoter we constructed multiple alignments using orthologous sequences from mouse, human, rhesus macaque, dog, cow, horse, and opossum, of the proximal promoter region consisting of 500 bps both upstream and downstream of the TSS (Balwierz et al., 2009). Using databases of experimentally determined binding sites (Wingender et al., 1996; Vlieghe et al., 2006), we collected a set of 207 mammalian regulatory motifs (position specific weight matrices) representing the binding specificities of approximately 350 mammalian TFs. Using a Bayesian probabilistic method that explicitly models the evolution of TFBSs, we then predict binding sites for all regulatory motifs in all proximal promoter regions (van Nimwegen, 2007). We summarise the binding site predictions by a matrix with components N_{pm}, denoting the sum of the posterior probabilities of all binding sites for motif m in promoter p, which we also refer to as the "number" of binding sites for motif m in promoter p. The second key ingredient of Epi-MARA is the quantification of epigenetic mark levels across the time course at genomic regions of interest. For the analysis of the ChIP-chip data, which measured H3K27me3 levels at all promoters genome-wide, we quantified the H3K7me3 at a given promoter and time point by the average log-intensity of the probes that lie within the promoter. For the ChIP-seq analysis we determined H3K27me3 enriched regions (see section 4.8) and found that the majority of H3K27me3 enriched regions are between 3 kb and 5 kb in length. For the analysis of ChIP-seq H3K27me3 levels at promoters we quantify the occurrence of H3K27me3 by the log-fraction of ChIP-seq reads in a 4kb region centred on the promoter. For the Epi-MARA analysis of genome-wide H3K27me3 enriched regions we use the log-fraction of ChIP-seq reads in each region.

Note that Epi-MARA fits the changes of H3K27me3 levels across the time course, thus Epi-MARA results are invariant to an overall rescaling of H3K27me3 levels at each promoter. Finally, to avoid spurious fluctuations in relative H3K27me3 levels at promoters with low absolute levels, a pseudo read-count corresponding the average read-count in the background sample is added to the read count in each promoter region. We denote the occurrence of the epigenetic mark M in promoter p at time point t by M_{pt} and assume the following linear

model:

$$M_{pt} = noise + c_p + \sum_m (N_{pm} * A_{ms}) \tag{4.1}$$

where c_p is the basal level of the chromatin mark, and A_{mt} is the unknown activity of motif m at time point t. Using a Bayesian probabilistic frame-work, we then calculate a joint posterior probability distribution for all motif activities. To this end, we assume that the deviation between model and measured level M_{pt} (i.e. the "noise" term in the above formula) is Gaussian distributed at each promoter and at each time point. In addition, to avoid over-fitting, we use a Gaussian prior on the activities A_{mt}, and we determine the variance of this prior by a cross-validation procedure. Finally, we infer both the maximal posterior activities and their standard-errors . To rank motifs, we measure the importance of a motif in explaining expression variations by a score similar to a z-statistic. The z-score z_m of motif m is quantified as an average squared z-value of the activity across conditions, i.e.

$$z_m = \sqrt{\frac{\sum_t (\frac{A^*_{mt}}{\sigma_{mt}})^2}{T}} \tag{4.2}$$

, where T is the number of time points. Note that our z-scores are meant to rank the importance of motifs and cannot be used to assess the statistical significance of motif activities. To run Epi-MARA on all H3K27me3 enriched regions genome-wide we predicted TFBSs across the entire 4 kb of each H3K27me3 region using the same procedure as used for proximal promoters. We then determined the 1 kb window that contains the highest number of predicted TFBS (pooling all motifs) and used these predicted sites for the entries in the site-count matrix N_{pm} for the corresponding H3K27me3 region. To infer motif activities separately for high- and low-CpG regions, for each motif m, we separately treated sites within low- and high-CpG regions as if they were derived from two separate motifs, effectively doubling the number of motifs for which we infer activities.

4.2 Cell culture and experimental system

Wildtype mouse embryonic stem cells were derived from blastocysts (3.5 PC) of mixed 129-C57Bl/6 background and cultivated on feeder cells (37C, 7% CO2). REST knock-out and corresponding wildtype cells were obtained from Helle Jorgensen (Chen et al., 1998; Jorgensen et al., 2009). Differentiation of cells was performed as described previously (Mohn et al., 2008; Bibel et al., 2007) taking advantage of a robust differentiation model for neurogenesis. ES cells are first differentiated into a highly homogeneous population of Pax6-positive radial-glial neuronal progenitor cells and further into terminally differentiated glutamatergic pyramidal

neurons (Bibel et al., 2004, 2007). In addition, transplantation experiments using chick embryos showed that neuronal progenitors are developmentally restricted to certain subtypes *in vivo* (Plachta et al., 2004). Terminally differentiated glutamatergic neurons, were further shown to possess defined electrophysiological characteristics resembling cortical glutamatergic neurons (Bibel et al., 2004, 2007).

4.3 Western blot analysis

For detection of REST protein levels during differentiation the total cell lysates of wildtype and REST knockout cells were used for western blot analysis. The membrane was probed with mouse anti-REST (12C11, gift from David Anderson) and rat anti-tubulin (tissue culture supernatant, cell line YL1/2, ECACC) in combination with appropriate secondary antibodies coupled to HRP.

4.4 Immunocytochemistry

Cells were fixed with 2% paraformaldehyde, either three hours or 10 days after plating, and probed with mouse anti-Pax6 (chick PAX6 a.a 1-223, DSHB), rabbit anti-Nestin (Sigma N5413) and mouse anti-Tuj1 (MMS-435P, Covance). Proteins were detected by an appropriate secondary antibody conjugated to Alexa Flour.

4.5 Chromatin immunoprecipitation (ChIP)

Cells were cross-linked in medium containing 1% formaldehyde for 10 min at room temperature. ChIP was carried out as previously described (Weber et al., 2007; Koch et al., 2007) with slight modifications. Antibodies used were α-H3K27me3 (Millipore,07-449) and α-REST (Santa Cruz, H-290), α-H3K4me2 (Millipore, 07-030), α-H3K4me3 (Millipore, 17-614) . Chromatin was sonicated for 15 (stem cells) or 18 cycles (neuronal progenitors) of 30 sec using a Diagenode Bioruptor, with 45 sec breaks in between. Precipitated DNA was either analysed by quantitative real time PCR or subjected to next generation sequencing.

4.6 Quantitative real-time PCR

Real time PCR was performed using SYBR green (ABI). 1/40 of ChIP sample or 40 ng of input chromatin were used per PCR reaction. Primer sequences are available upon request. All data is shown as either standard deviation or standard error from three biological replicates. All significances were calculated using unpaired 1-tailed student's t-test statistics.

4.7 Next generation sequencing

5 to 10 ng of precipitated DNA was prepared for Solexa Sequencing as described (Mikkelsen et al., 2007). Briefly, ChIP DNA was ligated to adapters and ligation products of about 250 bps were gel purified on 1.5 % agarose to remove unligated adaptors. DNA was amplified by 18 PCR cycles. DNA sequencing was carried out using the Illumina/Solexa Genome Analyzer II sequencing system. All generated data sets are available for download at the GEO database under GSE25533 and using the following URLs:
`www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=jxchzqgousyyeny&acc=GSE27148` `www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=vzqbzucsasqwgpq&acc=GSE27114`

Genomic coordinates: The July 2007 M. musculus genome assembly (NCBI37/mm9) provided by NCBI (`http://www.ncbi.nlm.nih.gov/genome/guide/mouse/`) and the Mouse Genome Sequencing Consortium (`http://www.sanger.ac.uk/Projects/M_musculus/`) was used as a basis for all analyses. Annotation of known RefSeq transcripts was obtained from UCSC (`http://hgdownload.cse.ucsc.edu/goldenPath/mm9/database/refGene.txt.gz`).

Read filtering, alignment and weighting low-complexity reads were filtered out based on their dinucleotide entropy as follows: For each read, the dinucleotide entropy was calculated according to the formula $H = \sum_i f_i log(f_i)$, where f_i is the frequency of dinucleotide i in the read and the sum is over all dinucleotides (i from 1 to 16). The read was filtered out if H was less than half the dinucleotide entropy of the genome, typically removing less than 0.5% of the reads in a given sample. Alignments to the mouse genome were performed by the software bowtie (version 0.9.9.1) (Langmead et al., 2009) with parameters -v 2 -a -m 100, tracking up to 100 best alignment positions per query and allowing at most two mismatches. Each alignment was weighted by the inverse of the number of hits. In the cases where a read had more hits to an individual sequence from the annotation database than to the whole genome, the former number of hits was selected to ensure that the total weight of a read does not exceed one. All quantification was based on weighted alignments. For generation

of wiggle files samples were normalised for library size first and files were generated with a window size of 100 bps.

4.8 Analysis of sequencing data

4.8.1 Identification of enriched regions

In order to detect REST peaks and H3K27me3 regions in the ChIP-seq data we pool all replicates and slide a 1 kb window for REST and a 2 kb window for H3K27me3 along the genome. We then calculate, for each window, the fraction f_{IP} of ChIP-seq reads from the IP and the fraction f_{bg} of reads from a background sample that map to this window (since background counts are generally smaller, we use a 2 kb window centred at the same position to obtain more robust background frequencies). Inspecting the reverse-cumulative distributions of background counts across the genome, we observed that a small subset of windows showed aberrantly high background frequencies f_{bg} (Figure 4.1) and removed these windows from further consideration (these windows are typically regions with repeats that presumably occur more frequently in the genome of the cells from which our DNA was taken, than in the mm9 genome assembly). We assume that the noise in the estimated f_{IP} and f_{bg} follow Poisson distributions and calculate, for each window, a z-statistic: $z = \frac{f_{IP} - f_{bg}}{\sqrt{\frac{f_{IP}}{N_{IP}} + \frac{f_{bg}}{N_{bg}}}}$, where N_{IP} and N_{bg} are the total numbers of reads in the IP and background sample, respectively. Inspecting the reverse-cumulative distribution of z-statistics across the genome, we observe a long tail of highly enriched regions to the right of z=3.1 for REST regions and z=4.0 for H3K27me3 regions (Figure 4.1) and set these values as cut-offs. All regions with z-values larger than the cut-off are defined as binding regions. This results in 1.618 REST binding regions and 18.293 H3K27me3 regions. Further REST and H3K27me3 enriched regions were divided into different classes using a number of criteria. Regions that overlapped the promoters (+- 2 kb around TSS) were considered proximal, all other regions were considered distal. Further, H3K27me3 regions that overlap a REST binding peak were considered REST targets and all other regions non-targets. We further calculated the CpG dinucleotide frequency within each window for each H3K27me3-enriched region by sliding a 1 kb window over the region. We defined the CpG-content of a region as the highest CpG frequency of a 1 kb window within it. Inspection of the distribution of log-CpG frequency across H3K27me3 enriched regions shows two classes that we fitted by a mixture of two Gaussians (Figure 3.5). After fitting of the Gaussian mixture, posterior probabilities for each region to belong to the high-CpG or low-CpG class were calculated in the standard Bayesian way. In subsequent analyses,

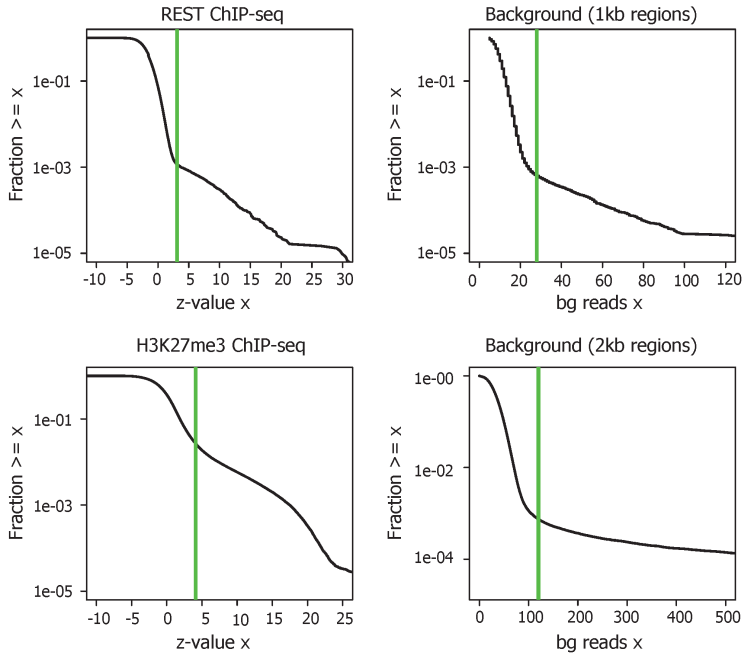

Figure 4.1: **left**) Reverse-cumulative distributions of the z-statistic for enrichment of ChIP-seq reads from REST (top) and H3K27me3 (bottom) IPs relative to background genome-wide. The distribution for REST shows two regimes with a second tail at z-statistics larger than approximately 3. The vertical line (z=3.1) shows the cut-off that we chose for considering a window significantly enriched for REST. For H3K27me3 two to three regimes are clearly evident in the distribution and we chose a cut-off of z=4.0 (vertical line) to identify H3K27me3-enriched windows. **right**) Reverse-cumulative distributions of background reads per 1 kb (top) or 2kb (bottom) windows genome-wide. For 1 kb regions the distribution drops steeply up to approximately 20 reads per window, after which it shows a long tail with some windows showing over 100 reads. We remove these genomic regions with high background counts (vertical line). For 2 kb regions the distribution similarly drops steeply up to approximately 100 reads per window. We again remove these genomic regions (vertical line).

distributions for low-CpG and high-CpG regions were obtained by weighing each region with the posterior probability that it belongs to the corresponding class.

4.8.2 REST binding site analysis

To predict REST binding sites for all REST binding regions we used two different approaches: We produced multiple alignments of orthologous regions from mouse, human, rhesus macaque, dog, cow, horse, and opossum, and ran the MotEvo algorithm (van Nimwegen, 2007; Arnold et al., 2012) on each multiple alignment. We also searched for non-canonical sites of arbitrary spacing between the two half-sites of the REST motif. As a second ap-

proach we also used MEME Suite (Bailey et al., 2009) searching for a 21 bp motif, with one occurrence per sequence and a 0-order background model. We then extended the identified motif by inserting a linker sequence from 1 to 11 nts in between the two REST half sites (Johnson et al., 2007). The linker sequences corresponded to background frequencies of the four bases and therefore did not influence the motif score. Using a cut-off of 1500 for the motif score we find 174.994 REST sites genome-wide, clearly predicting many false-positive sites. Linear regression between the total number of predicted REST sites (identified by MotEvo) at each REST binding region and the z-statistic of this region, shows a correlation of r=0.48 (P-value = $2.9 * 10^{-53}$). To obtain positional profiles with respect to TSS for the predicted binding sites of REST and other regulatory motifs, we summed the posteriors of all predicted sites at promoters at each position relative to TSS. To obtain positional profiles for the REST binding data and H3K27me3 signals we simply summed all reads from the corresponding IP samples at each position relative to TSS. To obtain positional profiles of H3K27me3 and SUZ12 relative to REST sites we selected all genomic regions that were enriched for H3K27me3 and overlapped a REST binding peak. For each of these we located the position of the highest scoring predicted REST binding site within the peak and then calculated the relative frequencies of reads, separately for SUZ12 and H3K27me3, in the 3 kb of sequence upstream and downstream of the REST site's position. We finally averaged these relative frequency profiles over all REST peaks within H3K27me3 enriched regions.

4.8.3 ChIP-seq data quantification

For each region that was enriched for H3K27me3 at any of the stages, we calculated log-fold changes between ES and NP and between NP and TN stages by calculating the log-ratios of the fractions of reads from the corresponding IP samples mapping to each of the regions. To compare H3K27me3 levels between wild type (WT) and RESTko (KO) cells we collected all regions that were enriched for H3K27me3 in the wild type cells at any of the stages. For each region we calculated the fractions f_{WT} and f_{KO} of all IP reads that mapped to that region in WT and KO and calculated both the absolute intensity X = ($\log(f_{WT})$ + $\log(f_{KO})$) (summed over all replicates) as well as the log-ratio: Y = $\log \frac{f_{WT}}{f_{KO}}$ (averaged over the replicates). Figure 4.2 shows, as a function of absolute intensity X, the average and standard error of Y for all regions that are non REST targets (black dots with error-bars) for high- and low-CpG regions at both the ES and NP stages. As this figure makes clear, there are some systematic differences in the overall distribution of H3K27me3 signals between

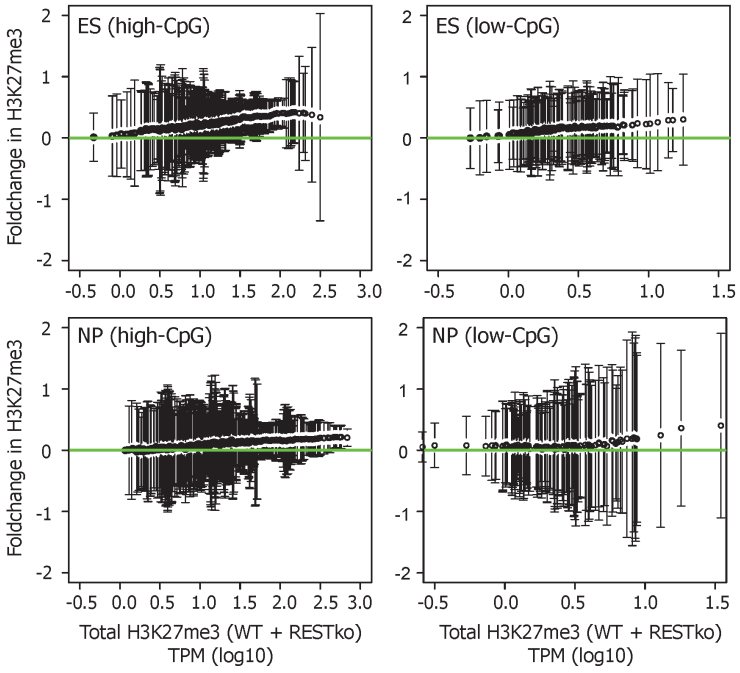

Figure 4.2: Shown are the total H3K27me3 levels (sum of wildtype and RESTko signal) (x-axis) against the average H3K27me3 fold-changes between wildtype and RESTko with standard errors (black dots with error-bars)(y-axis) for non REST target regions separated in high-CpG (**left**) and low-CpG (**right**) for both the ES (**top**) and NP (**bottom**) stages.

wildtype and the RESTko cells. In order to properly compare H3K27me3 signals between wildtype and RESTko, we therefore adopted a normalisation procedure similar to Loess normalisation: For each stage, we sorted all non-target regions by their absolute intensity X (averaging WT and KO intensities). For each region we then collected the 50 regions with values of X immediately below, and the 50 regions with values of X immediately above, and calculated the mean μ and standard deviation σ . In this way we estimated the expected mean μ and standard-deviation σ of non-targets, as a function of their absolute H3K27me3 levels. For each REST target we determined both its fold-change Y and absolute H3K27me3 level X and calculated a z-value $z = \frac{Y-\mu}{\sigma}$ using the expected mean and standard deviation of non-targets with absolute levels of H3K27me3 of X. To suppress fluctuations we averaged the z-statistics with a Gaussian kernel. Note that, per definition, the z-values of non-target regions follow a Gaussian distribution of mean zero and standard-deviation one. To estimate the fraction ρ of REST targets that significantly change H3K27me3 we compared the fraction of REST targets that show z-values more than one standard deviation away from the mean (i.e. z $>$ 1 when considering targets losing H3K27me3 and z $<$ -1 when considering targets gaining H3K27me3) with the fraction expected by chance using a Bayesian procedure. Let q

denote the probability to obtain a z-value larger than 1 by chance according to the standard Gaussian. Conservatively, assuming that all true targets must have a z-value larger than 1, the probability for a randomly chosen target to have a z-value larger than one is $p = \rho + (1-\rho)q$. Given that there are N REST targets in total, of which n have a z-value larger than 1 we use Bayes' theorem to calculate a posterior probability distribution over ρ and estimate its mean and standard-deviation. We similarly estimate the fraction of targets that significantly gain H3K27me3, separately for each stage, and separately for high- and low-CpG target regions. To compare SUZ12 levels in H3K27me3 enriched region of wildtype and RESTko neuronal progenitors, we determined the 1 kb region that had the highest overall read-count from the SUZ12 ChIP-seq. We then determined WT and KO SUZ12 levels from these 1kb regions. The z-statistics for the change in SUZ12 levels were calculated as above by comparing the log fold-change in SUZ12 of each REST target with those of the 100 non-target regions with the nearest absolute level in SUZ12 from the same CpG class.

4.8.4 Incorporating REST ChIP-seq data into Epi-MARA

To perform Epi-MARA analysis with the REST binding data replacing REST binding site predictions we replace the predicted binding site counts N_{pREST} with results of the REST binding assay at each promoter p. Since the z-statistics of REST binding at promoters have a very different distribution of values from those of the site counts N_{pm}, it is necessary to normalise the matrix N_{pm} such that binding site predictions and binding data can be quantitatively compared. We therefore replace the matrix N_{pm} with a binary matrix B_{pm} in which B_{pm} =1 whenever $N_{pm} > 0.2$ and $B_{pm} = 0$ otherwise. Finally, we replace the column B_{pREST} with one based on the REST binding data, i.e. where $B_{pREST} = 1$ whenever there was a REST binding peak within 2 kb of the corresponding promoter, and $B_{pREST} = 0$ otherwise.

4.9 RNA preparation and expression analysis

Total RNA was prepared using TRIzol (Invitrogen). mRNA expression data were generated using Mouse Gene 1.0 ST and Mouse Genome 430 2.0 arrays. Microarrays were RMA-normalised using R/Bioconductor (Gentleman et al., 2004) and the oligo package version 1.14.0 (Carvalho and Irizarry, 2010). To determine transcriptional regulation of REST target genes in the RESTko we selected a 2-fold change as cut-off for significant up-regulation.

4.10 Recombinase-mediated cassette exchange

Promoter fragments of REST or predicted SNAIL targets were cloned and stably integrated into stem cells via RMCE as described (Lienert et al., 2011b). Briefly, DNA fragments were cloned into a plasmid containing a multiple cloning site flanked by two inverted L1 Lox sites (gift from M. Lorincz). The cloned promoters varied in size between 1.2 and 2.5 kb. The following regions were clones: Stmn2 (chr3:8508035-8510167), Xkr7 (chr2:152857111-152858310), Bdnf (chr2:109514618-109516670), Pgbd5 (chr8:126957544-126958806), Cdh1 (chr8:109127058-109128426), Usp43 (chr11:67734626-67735819) and Esam (chr9:37335599-37338120). TC-1 ES cells were selected under hygromycin for 10 days, co-transfected with L1-promoter-1L plasmid and pIC-Cre and selected under ganciclovir. Clones were tested for successful insertion by PCR. ΔREST mutants were generated by removing 15 to 20 bps of the REST consensus sequence. The predicted SNAIL sites were mutated as described by changing two nucleotides of the core binding motif (Batlle et al., 2000).

Chapter 5

Discussion

Genome-wide analyses of chromatin have revealed unexpected dynamics of the epigenome, which reflect cellular and developmental states. The analysis of such data has predominantly focused on characterising the different kinds of chromatin domains that exist, and associating these domains with functional features such as active or inactive promoters or distal regulatory elements (Zhou et al., 2011; Suzuki and Bird, 2008; Meissner, 2010). With the exception of chromatin modifications that are set by the process of transcription itself, such as H3K36 methylation, our understanding of how dynamic changes in chromatin are regulated remains limited. This might reflect the complexity of the underlying targeting as different recruitment mechanisms for chromatin modifiers have been proposed (see section 1.3.3), including TFs, non-coding RNAs, as well as higher order nuclear organisation (Simon and Kingston, 2009; Beisel and Paro, 2011; Schuettengruber and Cavalli, 2009).

Notably, however, previous studies have mostly focused on systems in a single state, correlating the co-occurrence of a chromatin mark of interest with e.g. the frequency of DNA binding motifs of TFs, rather than inferring regulation based on temporal changes. Here, we have tested the hypothesis that TFs contribute to dynamic changes in chromatin during cellular differentiation. In collaboration with the van Nimwegen group, we combined mapping of epigenetic marks at consecutive stages with a computational methodology (Epi-MARA) that makes use of sophisticated binding site predictions for mammalian regulatory motifs to predict TFs involved in recruiting specific chromatin changes *ab initio*. We started from a data-set of mouse ES cells undergoing neurogenesis, in which levels of H3K27me3 were measured at three consecutive cellular states during the differentiation. Application of Epi-MARA to this data identified several TFs as potential regulators of Polycomb dynamics during differentiation (Figure 3.1). As a proof of concept we tested the prediction that

REST is involved in transiently recruiting H3K27me3 to promoter regions at the neuronal progenitor stage. We provide several lines of experimental evidence that support this model: 1. Genetic deletion shows that REST is necessary *in trans* for increased H3K27me3 levels at REST targets at the neuronal progenitor stage, specifically at high-CpG target regions (Figure 3.9), which includes almost all promoter-proximal target regions. Importantly, absence of REST further causes loss of the PRC2 component SUZ12, mirroring the loss H3K27me3 at high-CpG regions (Figure 3.12) and 3.13). 2. Promoter fragments containing TF binding sites of either REST or SNAIL TFs are sufficient *in cis* to recruit H3K27me3, whereas identical regions with mutated binding sites show reduced recruitment (Figure 3.16 and 3.17). 3. A substantial fraction of targets that lose H3K27me3 in the absence of REST show no corresponding transcriptional response (Figure 3.14), excluding a purely transcription-mediated effect. Together, these results show that local REST binding mediates Polycomb repression at the induction of *in vitro* neurogenesis.

As predicted by our computational model, REST has pronounced effects on H3K27me3 levels at target regions in neuronal progenitors. This context-dependent H3K27me3 recruitment might serve to repress neuronal genes, whose expression is only required at a later point of neuronal differentiation. This is compatible with the previous notion that REST function becomes more relevant after the neuronal fate decision, when precise regulation of neuronal genes is required (Chen et al., 1998; Jones and Meech, 1999).

While the detailed mechanisms of Polycomb targeting remain to be determined, our study provides several relevant insights. First, our results suggest that rather than a single dominating factor, targeting of Polycomb likely involves multiple TFs (see section 5.2). That is, we found that besides Rest several other regulatory motifs were associated with the transient increase of H3K27me3 at the neuronal progenitor stage, including Sp1, Snail, and Zeb1. ZEB1 and the family of SNAIL factors bind to very similar motifs and are important transcriptional repressors during differentiation processes such as EMT (Liu et al., 2008; Cano et al., 2000), which is compatible with a proposed function in Polycomb recruitment. Since Sp1 sites are among the most commonly occurring regulatory sites within CpG-islands, it is difficult to interpret whether the predicted role of Sp1 in H3K27me3 dynamics is specific to Sp1 or more generally associated with CpG-islands, which have been suggested to recruit PRC2 (Mendenhall et al., 2010). Of note, Sp1-like sites are a component of PREs in *Drosophila* (Brown and Kassis, 2010). In contrast, YY1, the mammalian ortholog of Pho, the most established TF with a function in Polycomb recruitment in *Drosophila melanogaster*, is unlikely to have that role in mammals (Mendenhall et al., 2010; Ku et al., 2008), at least in stem cells. Interestingly, like REST, the Snail and Zeb1 motifs also tend to be positioned immediately downstream of TSS (Figure 5.1). These findings are compatible with the model

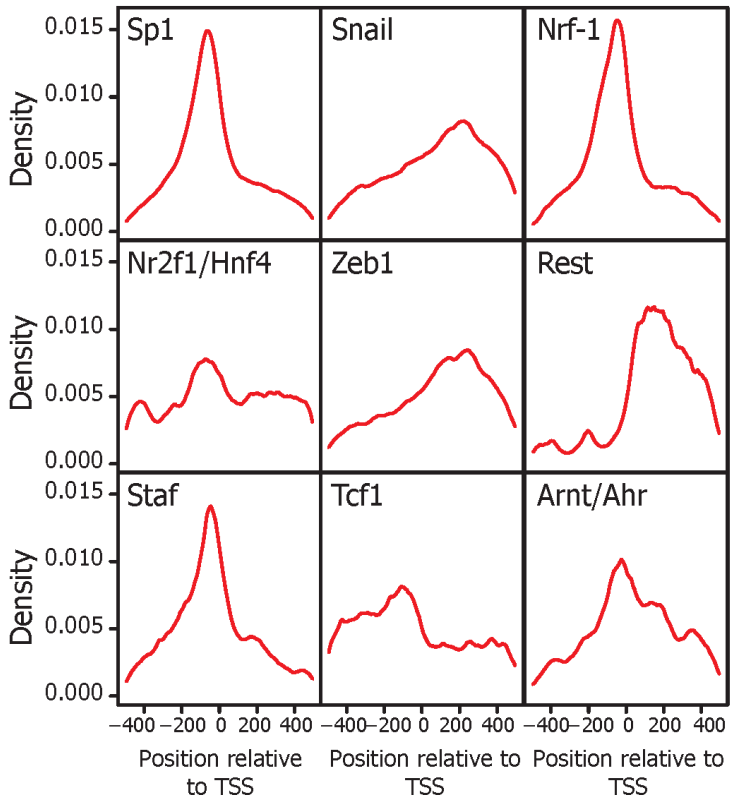

Figure 5.1: Frequencies of predicted binding sites around transcription start sites for Epi-MARA's top 9 predicted motifs. Sp1, Nrf-1, Staf and Arnt/Ahr show a strong binding preference around 50 bps upstream of TSS, while Snail, Zeb1, and Rest motifs are mostly found downstream of TSS. Nr2f1-Hnf4 and Tcf1 show much less pronounced positional preferences.

that Polycomb might repress by stalling polymerases (see discussion 5.2). Our results further show that the dynamics of H3K27me3 are generally distinct between high-CpG and low-CpG regions (Figure 3.7 and see discussion 5.2). Moreover, whereas REST has the strongest effect on H3K27me3 levels at high-CpG regions, it has a weaker opposite effect at low-CpG regions. These results suggest that either Polycomb is recruited through different mechanisms at high-CpG and low-CpG regions, or that the effects of local TF binding on H3K27me3 are modulated through a global mechanism with opposite action at low- and high-CpG regions (Mendenhall et al., 2010). REST binding decreases from the ES to NP stage, its effect on H3K27me3 however is by far strongest at the NP stage, indicating that stage-specific co-factors could mediate REST activity on Polycomb recruitment.

Epi-MARA provides a general methodology for predicting TFs associated with chromatin dynamics that will be highly useful for the study of epigenome maps. Key advantages of the approach are that it makes use of accurate TSS annotations to identify promoters combined

with sophisticated Bayesian comparative genomics methods to provide accurate predictions of TFBSs for a large number of mammalian regulatory motifs. Further, modelling dynamics of chromatin modifications across cellular stages in terms of TFBS, allows the specific identification of TFs that play a role in the changes of chromatin state, rather than identifying TFs that are generally enriched at regions bearing a particular chromatin mark. As Epi-MARA accounts for the contributions of all regulatory motifs at once, it is able to disentangle the contributions of different motifs that have correlated occurrences across promoters. The identification of REST exemplifies these advances. Its context-dependent role can only be uncovered by considering the dynamics of H3K27me3. Since Epi-MARA mainly depends on promoters and TFBS predictions that have been pre-calculated, the approach can easily be applied to any data-set measuring epigenome dynamics using various high-throughput measurement platforms. A multitude of such data sets are currently generated as part of large epigenome initiatives (Satterlee et al., 2010; Abbott, 2011) highlighting the need for methodologies that can go beyond classification of chromatin state patterns. The ability to predict TFs involved in regulating chromatin dynamics in a particular context from epigenome datasets provides a powerful tool in this context, as predicted TFs can be immediately subjected to follow-up experiments that further elucidate their regulatory role.

In the following sections I will discuss how the presented results relate to the role of Polycomb mediated repression(section 5.1), discuss evidence that argues for a role of TFs in the recruitment of histone modifications (section 5.2) and discuss the potential contribution of genetics to epigenetics (section 5.3). Finally, I will briefly discuss how binding specificity of TFs might me regulated (section 5.4).

5.1 Role of Polycomb-mediated repression

Polycomb-mediated gene repression is an essential component of the epigenetic machinery. In recent years many studies have documented the dynamic landscape of H3K27me3 levels and its cell-type specific targeting. In addition many PcG proteins are misregulated and in cancer, further stressing the relevance of this group of proteins. However, as pointed out in the introduction (1.3.3) the exact role of the Polycomb system is currently unclear as well as its mode of gene repression. This is also true for other chromatin modifications such as H3K4 methylation, where dynamics have been reported but the exact role of cell-type specific patterns remains obscure.

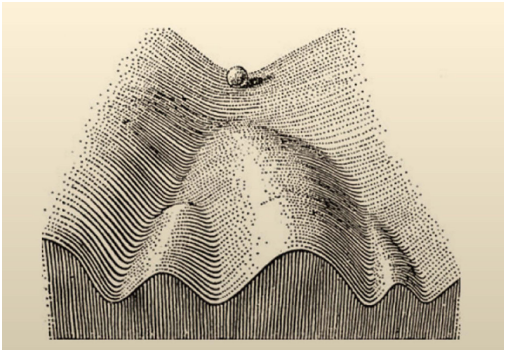

Figure 5.2: Conrad Waddingtons classical epigenetic landscape. The cell (represented by a ball) can only choose specific paths to distinct cell fates (visualised by trajectories down the hill) (Waddington, 1957)

The fact that many chromatin modifiers are essential for early development stresses the importance of chromatin dynamics in safeguarding transcriptional programs. Historically, the term "epigenetics" was coined by Conrad Waddington, who defined epigenetics as "the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being" (Waddington, 1942). In 1957 Waddington proposed his famous concept of an epigenetic landscape, where the process of cellular differentiation and commitment is guided by the environment (Waddington, 1957) (Figure 5.2). More specifically, the multi-potent cell (which is represented by a ball at the top of the hill) can take only certain permitted paths as it commits to a certain cell fate (and rolls down the hill). Even though this idea is over 50 years old it has lost nothing of its power. It is still a key concept that epigenetic modifications enhance robustness of transcriptional programs, reduce noise and define the developmental potential of cells as they differentiate. A recent report re-evaluated Waddington's concept of epigenetics and extended it by proposing that the amount of noise during differentiation is not homogeneous. Pluripotent cells would be noisier and have a landscape that is more flexible, whereas differentiated cells would be less noisy and less flexible (Pujadas and Feinberg, 2012). This model further assumes that noise should be highest during cell fate decisions in development. This suggests a role for epigenetics in noise reduction during cellular differentiation, which was suggested in particular for Polycomb-mediated

gene regulation (Mohn et al., 2008). The notion that epigenetic modifications repress noise and generate robustness is in agreement with our results. In our neurogenesis system we see a global increase of H3K27me3 at many high-CpG promoter-proximal regions from the transition of ES to NPs (Figure 3.7). In terms of cellular potential there is a large difference between ES cells and neuronal progenitors and it is plausible that Polycomb-mediated repression becomes more important as cells start to differentiate to a specific lineage. This effect is also visible in the RESTko cells. The transcriptional response in the absence of REST is the largest in neuronal progenitors, where most REST targets are transcriptionally up-regulated (Figure 3.8 c) arguing that REST-mediated repression is more important in NPs where the trans-acting environment might be more permissive of inappropriate up-regulation of neuronal genes.

It remains a key question how precisely Polycomb-mediated gene repression is functioning. Studies in *Drosophila melanogaster*, where Polycomb was first identified, defined PREs as genetic elements that act as recruiters of Polycomb and would confer active repression to a linked reporter gene (Cavalli and Paro, 1998; Ringrose and Paro, 2007). In mammals two human PREs have been identified that both were shown to repress genes (Woo et al., 2010; Sing et al., 2009). It is however likely that the ability of Polycomb to repress a reporter depends on the strength of that promoter. Therefore, it has not been comprehensively shown that Polycomb plays an active role in gene repression. Another scenario, in agreement with our findings, is that Polycomb acts downstream of the initial silencing by TFs to ensure that genes remain stably repressed.

Tethering PcG proteins to ectopic promoters can be very informative in this regard. Ideally, this experiment would be carried out with a library of different promoters to assess if the mediated repression by Polycomb is a function of promoter strength. To date only single gene studies have tried to address this question with contradictory results. The Helin lab tethered EED to a luciferase reporter that was driven by the *thymidine kinase* (TK) promoter. They detected a 2.5 fold reduction in luciferase activity, which was accompanied by an increase in H3K27me3 (Hansen et al., 2008). The Crabtree lab tethered EED to the OCT4 locus in mouse ES cells and fibroblasts. Importantly, this group showed that recruitment of Polycomb in ES cells did not induce repression of the reporter gene. In fibroblasts, where OCT4 is not expressed, the targeting of Polycomb was more effective. Moreover, the locus was shown to acquire H3K27me3 passively over several weeks (Hathaway and Bell et al., Cell, *in press*). This argues that Polycomb is not a specifier of gene silencing but rather secures the silenced state, that can be either actively set up by repressive TFs or passively by the absence of activating factors to safeguard transcriptional programs. Further experiments as suggested in section 5.5 will help to clarify the mechanism of Polycomb-mediated gene repression.

5.2 Role of TFs in recruitment of Polycomb

Transcription factors are an integral component of gene regulation. Even simple organism possess TFs that turn genes on and off (Browning and Busby, 2004) making TFBS highly predictive for gene expression (Beer and Tavazoie, 2004; Das et al., 2006; Hemberg and Kreiman, 2011; Irie et al., 2011). TFs existed before DNA was even packaged into nucleosomes, therefore it is likely that they interact with and possibly direct epigenetic pathways.
Most studies addressing Polycomb function in *Drosophila* are focusing on Hox loci, which represent the classical Polycomb targets. This class of genes is however very untypical compared to the average gene as Hox genes are extremely conserved and precisely regulated during development in a collinear fashion (Garcia-Fernandez, 2005). It is thus conceivable that Polycomb targeting might be different at Hox loci compared to other evolutionary younger developmental genes. The TF REST is an example of a rather young TF that is specific to vertebrates (see section 1.2.1). Its dynamic interaction with the evolutionary much older Polycomb system opens the possibility that TFs might generally utilise the epigenetic machinery, in particular Polycomb, to safeguard transcriptional programs.
A recent study showed that the PRC1 subunit Posterior sex combs (PSC) has a transcription-independent function in cell cycle progression. PSC is the *Drosophila* homologue of the mammalian PCGF and is a Ring domain protein that mediates mono-ubiquitination. Mohd-Sarip et al. showed that PSC interacts with the APC complex and ubiquitinates Cyclin B, which triggers its degradation and allows cell cycle progression (Mohd-Sarip et al., 2012). These data show that the enzymatic activity of PSC was exploited by a different protein complex suggesting that other Polycomb proteins might have been utilised by TFs during evolution.
We further predict multiple TFs to regulate targeting of Polycomb (Figure 3.1), which does not favour a model of direct interaction. The mechanism how TFs might recruit Polycomb is still unclear. In the case of REST previous studies have already noted increased H3K27me3 signal at REST-bound promoters and enrichment of REST binding sites at CpG-islands bound by PRC2 (Zheng et al., 2009; Ku et al., 2008). The absence of REST in stem cells has only subtle effects on H3K27me3 levels at target regions (Figure 3.9), suggesting that this previously noted co-occurrence of REST and H3K27me3 in stem cells has limited functional relevance. A more recent study showed that the non-coding RNA Hotair can bind to PRC2 and the LSD1/CoREST/REST complex *in vitro* (Tsai et al., 2010). However, Hotair is poorly conserved between human and mouse its absence in mouse has little effects on H3K27me3 levels *in vivo* (Schorderet and Duboule, 2011).

Motif	Correlation coefficient r	z-value (high-CpG)	z-value (low-CpG)
Negative correlation			
Ciz	-0.996	2.14	0.60
Rreb1	-0.995	2.12	2.87
Rest	-0.992	5.08	2.00
Ubx	-0.986	2.12	0.12
Gcnf	-0.932	2.23	1.01
Err	-0.905	2.15	0.45
Blimp1	-0.810	3.08	1.50
Weak correlation			
Sp1	-0.568	5.33	1.28
Irf-7	-0.540	2.06	0.11
Snail	-0.473	4.56	1.56
Tfap2a	-0.453	0.51	2.34
Foxd3	-0.295	3.80	2.17
Maz	-0.187	4.59	3.05
Sf-1	-0.085	3.56	0.49
Tef	-0.082	2.86	0.55
Prrx2	-0.078	2.77	4.08
Hmgiy	0.168	2.52	0.48
Myf	0.207	3.81	2.54
Gata-4	0.290	2.11	0.77
Sox5	0.591	0.73	2.00
Broad complex 4	0.592	2.62	0.54
Positive correlation			
Bzip Creb	0.696	2.46	0.16
FoxI1	0.782	2.44	1.92
Klf4	0.905	4.97	0.30
Fac1	0.952	2.49	0.01
Foxp1	0.984	1.21	2.88
Zeb1	0.995	3.52	0.47

Table 5.1: Shown are all motifs for which Epi-MARA predicted a z-value bigger than 2 (for either the high-CpG or low-CpG regions). For each motif the correlation coefficient r of the high-CpG and low-CpG profiles is shown in the second column. The z-scores for high-CpG and low-CpG activities are shown in the third and fourth columns. Whereas Rest shows a strong negative correlation, other TFs such as Zeb1 and Klf4 show a strong positive correlation. Thus, correlation coefficients are TF specific.

In addition, two recent studies report biochemical interaction between REST and members of the PRC1 and PRC2 complexes (Ren and Kerppola, 2011; Dietrich et al., 2012). Importantly, however these correlative observations at single cell states did not identify the dynamic and context-dependent role of REST on H3K27 methylation that we predict based on chromatin dynamics and further validate experimentally. Further, both studies identified

REST by mass spectrometry approaches, pointing out the limitation of such approaches to identify multiple TFs *ab initio* as we do using Epi-MARA.

In addition we infer TFs that are predicted to regulate H3K4me2 dynamics during *in vitro* neuronal differentiation (Figure 3.20) and experimentally test the prediction that REST regulates a loss of H3K4 methylation at the transition of ES cells to NPs (see section 3.7). We indeed detect an increase in H3K4me3 at promoter-proximal REST binding regions in RESTko neuronal progenitors. However, this gain is better explained by changes in gene expression than changes in H3K27me3 levels in the absence of REST (see equation3.2). However, H3K27 methylation changes still contribute significantly to the linear model, thus further experiments will be needed to dissect the interplay of H3K4 and H3K27 methylation and the impact of these modifications on gene expression. Especially causal relationships still need to be clarified. Experiments perturbing the expression levels of H3K27 and H3K4-specific methylases or demethylases will be highly informative regarding the potential interplay between theses two epigenetic pathways.

Based on recent work in *Drosophila melanogaster* (Enderle et al., 2010) and mouse stem cells (Landeira et al., 2010) it has been suggested that Polycomb might exert its repressive function by stalling polymerases. Our predicted binding site location of the TFs REST, SNAIL and ZEB1, that are all predicted to bind downstream of TSS, is compatible with that model (Figure 5.1). TF binding downstream of the TSS might form a physical barrier that interferes with the transcriptional machinery. Supporting this hypothesis, is a study in *Drosophila melanogaster* embryos, showing that the SNAIL TF inhibits release of Pol II from promoters (Bothma et al., 2011).

Interestingly, Epi-MARA predicts distinct motif activities for high- and low-CpG regions. Depending on the motif these high- and low-CpG activities can be correlated or not, arguing that the different activities are not due to a general sequence bias but rather specific for each TF (Table 5.1). We experimentally validate the specific role for REST at high- and low-CpG H3K27me3 regions genome-wide (Figure 3.9). Further studies will be necessary to uncover the role of REST at low-CpG regions. A recent study reported that REST binding is enriched at distal regulatory regions (Stadler et al., 2011) that overlap enhancers, suggesting that these distal REST binding sites are functional. However, the observed up-regulation of REST targets in RESTko neuronal progenitors can almost entirely be explained by promoter-proximal REST peaks arguing that the effect of distal REST peaks on gene expression is little. Therefore, further experiments, involving the mapping of enhancer-promoter interactions, will be needed to study the contribution of distal REST binding sites to gene regulation. Several studies have argued that CpG-density is a major contributor to H3K27me3 recruitment (Mendenhall et al., 2010; Ku et al., 2008) however as CpG-rich sequences are highly enriched

Observables	H3K27me3 levels (ChIP-chip)	Transcript levels (chip)
	8.0%	6.9%

Table 5.2: Comparison of observed expression (measured by micro-array) and H3K27me3 (measured by ChIP-chip) levels, with the levels predicted by MARA (Suzuki et al., 2009) and Epi-MARA. Using the linear model with the fitted motif activities A_{ms}, we calculate what fraction of the total variance (summed over all promoters) is explained by the linear model. A similar fraction of variance for both H3K27me3 and expression dynamics are explained across the differentiation.

for TFBS it is difficult to separate the contribution of TFBS and CpG content, which requires future experiment (see 5.5). Further suggesting a connection between Polycomb and DNA methylation is the finding that promoters that are bound by Polycomb in ES cells are much more likely to be de novo methylated upon differentiation compared to promoters that are not bound by Polycomb (Mohn et al., 2008). A possible explanation for this scenario could be that these regions are occupied by TFs that recruit Polycomb in ES cells. In NPs these factors could be absent leading to DNA methylation of these regions in agreement with recent studies showing that TF-binding leads to a local loss of DNA methylation (Stadler et al., 2011; Lienert et al., 2011b).

5.3 A role for genetics in epigenetics

The validation that TFs regulate Polycomb dynamics implies that epigenetic gene regulation is to some extent regulated by the genetic information encoded in the genome. We further show that promoter fragments containing TFBSs are sufficient to recruit Polycomb-mediated H3K27me3 to similar levels as the endogenous promoters (Figure 3.16 and 3.17) arguing that the required information for Polycomb targeting is locally encoded in the DNA sequence. Importantly, in our neurogenesis paradigm, a linear model in terms of predicted binding sites explains roughly the same fraction of variance in H3K27me3 at promoters as it explains variance in transcript levels (Table 5.2). This result further suggests that, like regulation of transcription, H3K27me3 dynamics are generally regulated by local DNA motifs that are recognised by trans-acting factors.

Further suggesting a central contribution of DNA sequence to epigenetics is the finding that TF binding is involved in the protection against DNA methylation (Macleod et al., 1994; Brandeis et al., 1994; Dickson et al., 2010). Moreover, TFs in general can lead to reduced DNA methylation at regulatory regions genome-wide (Stadler et al., 2011). In line with a function for DNA sequence, it was further shown that inserted promoter fragments are suffi-

cient to regulate stage-specific DNA methylation patterns compared to endogenous sequences, strongly arguing that DNA methylation is to a large extent regulated by local DNA sequence (Lienert et al., 2011b). It was already shown in 1979 that transcription creates DNAse I hypersensitive sites (Wu et al., 1979), arguing that TF binding is upstream of changes in chromatin structure (Degner et al., 2012). A classical definition of the word "epigenetic" is defined as "a change in the state of expression of a gene that does not involve a mutation, but that is nevertheless inherited in the absence of the signal (or event) that initiated that change" (Ptashne and Gann, 2002). This definition is compatible with the idea that DNA binding elements are responsible for setting up a specific chromatin state that is then stably retained even in the absence of the TF.

A potentially powerful method for dissecting the interplay between DNA sequence and expression or chromatin state are quantitative trait loci (QTLs) (Gilad et al., 2008). Expression QTLs (eQTLs) associate genetic polymorphisms, such as single nucleotide polymorphisms (SNPs), with gene expression changes. Recent studies showed that eQTLs as well as DNase I sensitivity QTLs (dsQTLs) frequently enrich for TFBSs (Gaffney et al., 2012; Degner et al., 2012). Furthermore, dsQTLs where shown to be frequently associated with allele-specific changes in TF binding. This data argue that QTL are likely important contributors to phenotypic variation.

Further studies will show to what extent DNA sequence determines gene expression and epigenetics. Importantly, this implies that epigenetic changes between individuals could be explained by genetic heterogeneity. Therefore, studies linking epigenetic changes to certain phenotypes would have to take differences in DNA sequence into account.

5.4 How is TF binding regulated?

The presented data argues for a TF-based regulation of chromatin dynamics. The linear model applied in the Epi-MARA approach can explain about 8 % of the variance observed for H327me3 dynamics. This fraction indicates a highly significant contribution of TFs in explaining dynamics of Polycomb, however, 92 % of the variance remain unexplained. It is tempting to speculate how this contribution would change if more accurate binding maps of TFs were available. This will on the one hand require further characterisation of TFs for which no binding specificity is currently known, and which still represent the majority of mammalian TFs. On the other hand it is crucial to generate more binding maps of TFs *in vivo* to analyse if *in vitro* determined TF motifs can generally predict binding sites *in vivo*. It is at this point not clear if more comprehensive TF binding maps will increase the

explained fraction of changes in gene expression and chromatin dynamics but it seems likely to be the case. Even in the case of REST, where the binding motif is untypically long, the incorporation of actual binding data into Epi-MARA still doubled the significance of the prediction (Figure 3.4).

Another question directly related to further characterising TF binding is aimed at understanding how TF binding specificity is regulated and if specificity is influenced by the chromatin state of target loci. The analysis of dynamic REST binding during neuronal differentiation showed that REST motif quality plays a key role in retaining REST at high affinity sites when REST protein levels decrease (Figure 3.25). It is currently unclear if this is generally the case for TFs. In particular, TFs with short consensus sequences such as the SNAIL family of TFs (see 1.2.2) have a large number of perfect binding motifs in the genome. Thus it is plausible that additional mechanisms regulate specific binding. In this regard histone modifications that can regulate accessibility of DNA might discriminate functional and non-functional binding sites. Further sequential or cooperative binding of TFs might be required to create functional binding sites.

The increase in genome size during evolution of more complex eukaryotic species was accompanied by novel mechanisms to regulate transcriptional noise (Bird, 1995). Many studies showed that the packaging of DNA in form of nucleosomes decreases *in vitro* transcription (Felsenfeld, 1992; Lorch et al., 1987). Thus, histone modifications are likely to have evolved to regulate accessibility of the DNA template (Mohn and Schubeler, 2009). Supporting this notion are several studies presenting evidence that chromatin shapes TF binding patterns (Guccione et al., 2006; John et al., 2011). If this is generally the case is subject of further studies. Another study showed that master TFs, such as OCT4 and MYOD1 can act as "pioneering" factors and direct SMAD3 binding to DNA arguing for a sequential recruitment of TFs. Interestingly, this study further showed that SMAD3 can be broadly redirected by the overexpression of a master TF from another cell-type (Mullen et al., 2011). Pioneering factors are often lineage-specific and thought to have the ability to bind to nucleosome-wrapped DNA leading to the subsequent recruitment of chromatin remodelling complexes and the establishment of low nucleosome occupancy regions that allow other TFs to bind (Zaret and Carroll, 2011). Another study in yeast came to different conclusions showing that binding of the TF MSN2P was required for the loss of nucleosomes (Huebert et al., 2012). It is likely that the influence of chromatin on TF binding is TF-specific and possibly depends on the redundancy of the binding motif or other parameters. Future studies that measure TF binding dynamics together with a measurement for chromatin accessibility will likely be insightful to answer these questions.

Cooperative binding of TFs is also likely to enhance stable and thus functional TF binding

sites. Analysis of TWIST binding in *Drosophila* embryos suggested that functional TWIST binding sites partially depend on neighbouring SNAIL or DORSAL binding sites, which is supported by a very similar binding pattern of these three TFs (Zeitlinger et al., 2007; He et al., 2011). Similarly, core members of the pluripotency network in stem cells often cooperate in their binding to DNA. OCT4, NANOG and SOX2 share a large subset of genomic targets (Young, 2011; Chen et al., 2008; Boyer et al., 2005).

5.5 Further experiments and outlook

A key remaining question is the order of events of Polycomb mediated repression. To further dissect the temporal relationship between H3K27me3 and gene expression an interesting experiment would be to re-introduce REST in the RESTko cells. In the absence of REST several REST bound promoter regions lose H3K27me3 and a subset of these is transcriptionally up-regulated. Reintroducing REST into ES cells in a high resolution time course could address the question if transcriptional silencing would occur before re-establishment of correct H3K27me3 levels. These experiments are limited by the ability of the respective antibodies to detect low levels of a modification, but this experiment will nevertheless be informative about the order of events that take place to repress genes both by transcriptional and epigenetic means. Similarly, one could further reintroduce REST in the terminally differentiated neurons. Here, most REST targets are expressed and it would be interesting to assess if H3K27me3 is reestablished once REST targets are repressed again.
To further address the mechanism of Polycomb recruitment one could interfere with the expression of activating TFs. If the Polycomb machinery acts generally downstream of transcriptional repression it is likely that the absence of a certain activating TF, which would result in down-regulation of a specific set of genes, is accompanied by an increase in H3K27me3 at these genes. It is feasible that the repression of Polycomb depends on the strength of the linked promoter. Most studies addressing Polycomb-mediated gene repression in mammals have used a similar promoter with different upstream Polycomb recruitment fragment (Hansen et al., 2008). It would be interesting to instead take a fixed Polycomb recruiter fragment but fuse it to a variety of different promoters, to analyse the extent of Polycomb-mediated repression depending on promoter strength.
To further dissect the contribution of TFBS and CpG content on Polycomb recruitment several experiments could be performed. For once one could take a CpG-rich promoter sequence that recruits H3K27me3 and randomise the entire sequence except the CpGs. This way the contribution of potential TFBSs and CpG content could be uncoupled. As an opposing ap-

proach one could insert REST sites into CpG Islands that do not harbour H3K27me3 and test whether REST binding leads to an increase in H3K27me3. Another possibility to test the contribution of CpGs would be to use bacterial DNA, that should *per se* be depleted of mammalian TFBS. In a high-throughput manner it could be assessed whether the recruitment of H3K27me3 depends on the CpG content of the bacterial DNA.
Together, a lot of questions still remain. However, sophisticated library approaches, coupled to high-throughput imaging and sequencing technologies in combination with advanced data analysis will make it feasible to uncover the contribution of genetics to epigenetics and the role of epigenetic modifications for gene regulation.

Chapter 6

References

A. Abbott. Europe to map the human epigenome. *Nature*, 477(7366):518, 2011.

A. Alberga, J. L. Boulay, E. Kempe, C. Dennefeld, and M. Haenlin. The snail gene required for mesoderm formation in drosophila is expressed dynamically in derivatives of all three germ layers. *Development*, 111(4):983–92, 1991.

B. Alberts, A. Johnson, and J. Lewis. *Molecular Biology of the Cell*. Chromosomal DNA and Its Packaging in the Chromatin Fiber. Garland Science, New York, 4th edition, 2002.

L. Alland, R. Muhle, Jr. Hou, H., J. Potes, L. Chin, N. Schreiber-Agus, and R. A. DePinho. Role for n-cor and histone deacetylase in sin3-mediated transcriptional repression. *Nature*, 387(6628):49–55, 1997.

M. E. Andres, C. Burger, M. J. Peral-Rubio, E. Battaglioli, M. E. Anderson, J. Grimes, J. Dallman, N. Ballas, and G. Mandel. Corest: a functional corepressor required for regulation of neural-specific gene expression. *Proc Natl Acad Sci U S A*, 96(17):9873–8, 1999.

K. J. Armache, J. D. Garlick, D. Canzio, G. J. Narlikar, and R. E. Kingston. Structural basis of silencing: Sir3 bah domain in complex with a nucleosome at 3.0 a resolution. *Science*, 334(6058):977–82, 2011.

P. Arnold. *A Framework to Identify Epigenome and Transcription Factor Crosstalk*. PhD thesis, University of Basel, 2011.

P. Arnold, I. Erb, M. Pachkov, N. Molina, and E. van Nimwegen. Motevo: integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of dna sequences. *Bioinformatics*, 28(4):487–94, 2012.

T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. Meme suite: tools for motif discovery and searching. *Nucleic Acids Res*, 37(Web Server issue):W202–8, 2009.

N. Ballas and G. Mandel. The many faces of rest oversee epigenetic programming of neuronal genes. *Curr Opin Neurobiol*, 15(5):500–6, 2005.

D. Baltimore. Our genome unveiled. *Nature*, 409(6822):814–6, 2001.

P. J. Balwierz, P. Carninci, C. O. Daub, J. Kawai, Y. Hayashizaki, W. Van Belle, C. Beisel, and E. van Nimwegen. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepcage data. *Genome Biol*, 10(7):R79, 2009.

A. J. Bannister, P. Zegerman, J. F. Partridge, E. A. Miska, J. O. Thomas, R. C. Allshire, and T. Kouzarides. Selective recognition of methylated lysine 9 on histone h3 by the hp1 chromo domain. *Nature*, 410(6824):120–4, 2001.

A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37, 2007.

E. Batlle, E. Sancho, C. Franci, D. Dominguez, M. Monfar, J. Baulida, and A. Garcia De Herreros. The transcription factor snail is a repressor of e-cadherin gene expression in epithelial tumour cells. *Nat Cell Biol*, 2(2):84–9, 2000.

M. A. Beer and S. Tavazoie. Predicting gene expression from sequence. *Cell*, 117(2):185–98, 2004.

C. Beisel and R. Paro. Silencing chromatin: comparing modes and mechanisms. *Nat Rev Genet*, 12(2):123–35, 2011.

O. Bell, M. Schwaiger, E. J. Oakeley, F. Lienert, C. Beisel, M. B. Stadler, and D. Schubeler. Accessibility of the drosophila genome discriminates pcg repression, h4k16 acetylation and replication timing. *Nat Struct Mol Biol*, 17(7):894–900, 2010.

O. Bell, V. K. Tiwari, N. H. Thoma, and D. Schubeler. Determinants and dynamics of genome accessibility. *Nat Rev Genet*, 12(8):554–64, 2011.

C. M. Bergman, J. W. Carlson, and S. E. Celniker. Drosophila dnase i footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, drosophila melanogaster. *Bioinformatics*, 21(8):1747–9, 2005.

B. E. Bernstein, T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, and E. S. Lander. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–26, 2006a.

B. E. Bernstein, A. Meissner, and E. S. Lander. The mammalian epigenome. *Cell*, 128(4): 669–81, 2007.

E. Bernstein, E. M. Duncan, O. Masui, J. Gil, E. Heard, and C. D. Allis. Mouse polycomb proteins bind differentially to methylated histone h3 and rna and are enriched in facultative heterochromatin. *Mol Cell Biol*, 26(7):2560–9, 2006b.

M. Bibel, J. Richter, K. Schrenk, K. L. Tucker, V. Staiger, M. Korte, M. Goetz, and Y. A. Barde. Differentiation of mouse embryonic stem cells into a defined neuronal lineage. *Nat Neurosci*, 7(9):1003–9, 2004.

M. Bibel, J. Richter, E. Lacroix, and Y. A. Barde. Generation of a defined and uniform population of cns progenitors and neurons from mouse embryonic stem cells. *Nat Protoc*, 2(5):1034–43, 2007.

Y. Bilu and N. Barkai. The design of transcription-factor binding sites is affected by combinatorial regulation. *Genome Biol*, 6(12):R103, 2005.

A. Bird. Dna methylation patterns and epigenetic memory. *Genes Dev*, 16(1):6–21, 2002.

A. P. Bird. Cpg-rich islands and the function of dna methylation. *Nature*, 321(6067):209–13, 1986.

A. P. Bird. Gene number, noise reduction and biological complexity. *Trends Genet*, 11(3): 94–100, 1995.

A. P. Bird and A. P. Wolffe. Methylation-induced repression–belts, braces, and chromatin. *Cell*, 99(5):451–4, 1999.

E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, and P. J. de Jong. Identification and analysis of functional elements in 1genome by the encode pilot project. *Nature*, 447(7146):799–816, 2007.

J. P. Bothma, J. Magliocco, and M. Levine. The snail repressor inhibits release, not elongation, of paused pol ii in the drosophila embryo. *Curr Biol*, 21(18):1571–7, 2011.

D. Bourc'his and T. H. Bestor. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking dnmt3l. *Nature*, 431(7004):96–9, 2004.

P. Bouwman and S. Philipsen. Regulation of the activity of sp1-related transcription factors. *Mol Cell Endocrinol*, 195(1-2):27–38, 2002.

L. A. Boyer, T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch, and

R. A. Young. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6):947–56, 2005.

L. A. Boyer, K. Plath, J. Zeitlinger, T. Brambrink, L. A. Medeiros, T. I. Lee, S. S. Levine, M. Wernig, A. Tajonar, and R. Jaenisch. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, 441(7091):349–53, 2006.

A. P. Bracken, N. Dietrich, D. Pasini, K. H. Hansen, and K. Helin. Genome-wide mapping of polycomb target genes unravels their roles in cell fate transitions. *Genes Dev*, 20(9): 1123–36, 2006.

M. Brandeis, D. Frank, I. Keshet, Z. Siegfried, M. Mendelsohn, A. Nemes, V. Temper, A. Razin, and H. Cedar. Sp1 elements protect a cpg island from de novo methylation. *Nature*, 371(6496):435–8, 1994.

S. D. Briggs, T. Xiao, Z. W. Sun, J. A. Caldwell, J. Shabanowitz, D. F. Hunt, C. D. Allis, and B. D. Strahl. Gene silencing: trans-histone regulatory pathway in chromatin. *Nature*, 418(6897):498, 2002.

E. Brookes, I. de Santiago, D. Hebenstreit, K. J. Morris, T. Carroll, S. Q. Xie, J. K. Stock, M. Heidemann, D. Eick, and A. Pombo. Polycomb associates genome-wide with a specific rna polymerase ii variant, and regulates metabolic genes in escs. *Cell Stem Cell*, 10(2): 157–70, 2012.

J. L. Brown and J. A. Kassis. Spps, a drosophila sp1/klf family member, binds to pres and is required for pre activity late in development. *Development*, 137(15):2597–602, 2010.

D. F. Browning and S. J. Busby. The regulation of bacterial transcription initiation. *Nat Rev Microbiol*, 2(1):57–65, 2004.

A. W. Bruce, I. J. Donaldson, I. C. Wood, S. A. Yerbury, M. I. Sadowski, M. Chapman, B. Gottgens, and N. J. Buckley. Genome-wide analysis of repressor element 1 silencing transcription factor/neuron-restrictive silencing factor (rest/nrsf) target genes. *Proc Natl Acad Sci U S A*, 101(28):10458–63, 2004.

P. Bucher. Weight matrix descriptions of four eukaryotic rna polymerase ii promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol*, 212(4):563–78, 1990.

G. Buchwald, P. van der Stoop, O. Weichenrieder, A. Perrakis, M. van Lohuizen, and T. K. Sixma. Structure and e3-ligase activity of the ring-ring complex of polycomb proteins bmi1 and ring1b. *EMBO J*, 25(11):2465–74, 2006.

C. Buecker and J. Wysocka. Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet*, 2012.

M. Buhler, A. Verdel, and D. Moazed. Tethering rits to a nascent transcript initiates rnai- and heterochromatin-dependent gene silencing. *Cell*, 125(5):873–86, 2006.

T. Burgold, F. Spreafico, F. De Santa, M. G. Totaro, E. Prosperini, G. Natoli, and G. Testa. The histone h3 lysine 27-specific demethylase jmjd3 is required for neural commitment. *PLoS One*, 3(8):e3034, 2008.

C. D. Bustamante, A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz, S. Glanowski, D. M. Tanenbaum, T. J. White, J. J. Sninsky, and A. G. Clark. Natural selection on protein-coding genes in the human genome. *Nature*, 437(7062):1153–7, 2005.

A. Cano, M. A. Perez-Moreno, I. Rodrigo, A. Locascio, M. J. Blanco, M. G. del Barrio, F. Portillo, and M. A. Nieto. The transcription factor snail controls epithelial-mesenchymal transitions by repressing e-cadherin expression. *Nat Cell Biol*, 2(2):76–83, 2000.

R. Cao and Y. Zhang. Suz12 is required for both the histone methyltransferase activity and the silencing function of the eed-ezh2 complex. *Mol Cell*, 15(1):57–67, 2004.

R. Cao, L. Wang, H. Wang, L. Xia, H. Erdjument-Bromage, P. Tempst, R. S. Jones, and Y. Zhang. Role of histone h3 lysine 27 methylation in polycomb-group silencing. *Science*, 298(5595):1039–43, 2002.

T. F. Carl, C. Dufton, J. Hanken, and M. W. Klymkowsky. Inhibition of neural crest migration in xenopus using antisense slug rna. *Dev Biol*, 213(1):101–15, 1999.

P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engstrom, and Y. Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38(6):626–35, 2006.

S. B. Carroll. Chance and necessity: the evolution of morphological complexity and diversity. *Nature*, 409(6823):1102–9, 2001.

S. B. Carroll. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, 134(1):25–36, 2008.

M. J. Carrozza, B. Li, L. Florens, T. Suganuma, S. K. Swanson, K. K. Lee, W. J. Shia, S. Anderson, J. Yates, and J. L. Workman. Histone h3 methylation by set2 directs deacetylation of coding regions by rpd3s to suppress spurious intragenic transcription. *Cell*, 123(4): 581–92, 2005.

B. S. Carvalho and R. A. Irizarry. A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, 26(19):2363–7, 2010.

E. A. Carver, R. Jiang, Y. Lan, K. F. Oram, and T. Gridley. The mouse snail gene encodes a key regulator of the epithelial-mesenchymal transition. *Mol Cell Biol*, 21(23):8184–8, 2001.

G. Cavalli and R. Paro. The drosophila fab-7 chromosomal element conveys epigenetic inheritance during mitosis and meiosis. *Cell*, 93(4):505–18, 1998.

S. C. Chang, T. Tucker, N. P. Thorogood, and C. J. Brown. Mechanisms of x-chromosome inactivation. *Front Biosci*, 11:852–66, 2006.

W. T. Chang, H. I. Chen, R. J. Chiou, C. Y. Chen, and A. M. Huang. A novel function of transcription factor alpha-pal/nrf-1: increasing neurite outgrowth. *Biochem Biophys Res Commun*, 334(1):199–206, 2005.

K. Chen and N. Rajewsky. The evolution of gene regulation by transcription factors and micrornas. *Nat Rev Genet*, 8(2):93–103, 2007.

X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, and H. H. Ng. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–17, 2008.

Z. F. Chen, A. J. Paquette, and D. J. Anderson. Nrsf/rest is required in vivo for repression of multiple neuronal target genes during embryogenesis. *Nat Genet*, 20(2):136–42, 1998.

J. M. Claverie. Gene number. what if there are only 30,000 human genes? *Science*, 291 (5507):1255–7, 2001.

G.M. Cooper. *The Cell: A Molecular Approach*. Sinauer Associates, Sunderland (MA), 2nd edition, 2000.

B. Czermin, R. Melfi, D. McCabe, V. Seitz, A. Imhof, and V. Pirrotta. Drosophila enhancer of zeste/esc complexes have a histone h3 methyltransferase activity that marks chromosomal polycomb sites. *Cell*, 111(2):185–96, 2002.

J. Dai, E. M. Hyland, D. S. Yuan, H. Huang, J. S. Bader, and J. D. Boeke. Probing nucleosome function: a highly versatile library of synthetic histone h3 and h4 mutants. *Cell*, 134(6):1066–78, 2008.

D. Das, Z. Nahle, and M. Q. Zhang. Adaptively inferring human transcriptional subnetworks. *Mol Syst Biol*, 2:2006 0029, 2006.

M. de Hoon and Y. Hayashizaki. Deep cap analysis gene expression (cage): genome-wide identification of promoters, quantification of their expression, and network inference. *Biotechniques*, 44(5):627–8, 630, 632, 2008.

M. de Napoles, J. E. Mermoud, R. Wakao, Y. A. Tang, M. Endoh, R. Appanah, T. B. Nesterova, J. Silva, A. P. Otte, M. Vidal, and N. Brockdorff. Polycomb group proteins ring1a/b link ubiquitylation of histone h2a to heritable gene silencing and x inactivation. *Dev Cell*, 7(5):663–76, 2004.

J. F. Degner, A. A. Pai, R. Pique-Regi, J. B. Veyrieras, D. J. Gaffney, J. K. Pickrell, S. De Leon, K. Michelini, N. Lewellen, and J. K. Pritchard. Dnase i sensitivity qtls are a major determinant of human expression variation. *Nature*, 482(7385):390–4, 2012.

J. Dickson, H. Gowher, R. Strogantsev, M. Gaszner, A. Hair, G. Felsenfeld, and A. G. West. Vezf1 elements mediate protection from dna methylation. *PLoS Genet*, 6(1):e1000804, 2010.

N. Dietrich, M. Lerdrup, E. Landt, S. Agrawal-Singh, M. Bak, N. Tommerup, J. Rappsilber, E. Sodersten, and K. Hansen. Rest-mediated recruitment of polycomb repressor complexes in mammalian cells. *PLoS Genet*, 8(3):e1002494, 2012.

M. F. Dion, S. J. Altschuler, L. F. Wu, and O. J. Rando. Genomic characterization reveals a simple histone h4 acetylation code. *Proc Natl Acad Sci U S A*, 102(15):5501–6, 2005.

D. M. Eisenmann, C. Dollard, and F. Winston. Spt15, the gene encoding the yeast tata binding factor tfiid, is required for normal transcription initiation in vivo. *Cell*, 58(6): 1183–91, 1989.

G. Elgar and T. Vavouri. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet*, 24(7):344–52, 2008.

S. C. Elgin. The formation and function of dnase i hypersensitive sites in the process of gene activation. *J Biol Chem*, 263(36):19259–62, 1988.

D. Enderle, C. Beisel, M. B. Stadler, M. Gerstung, P. Athri, and R. Paro. Polycomb preferentially targets stalled promoters of coding and noncoding transcripts. *Genome Res*, 21 (2):216–26, 2010.

S. Erhardt, I. H. Su, R. Schneider, S. Barton, A. J. Bannister, L. Perez-Burgos, T. Jenuwein, T. Kouzarides, A. Tarakhovsky, and M. A. Surani. Consequences of the depletion of zygotic and embryonic enhancer of zeste 2 during preimplantation mouse development. *Development*, 130(18):4235–48, 2003.

R. Eskeland, M. Leeb, G. R. Grimes, C. Kress, S. Boyle, D. Sproul, N. Gilbert, Y. Fan, A. I. Skoultchi, and W. A. Bickmore. Ring1b compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Mol Cell*, 38(3):452–64, 2010.

C. Faust, A. Schumacher, B. Holdener, and T. Magnuson. The eed mutation disrupts anterior mesoderm production in mice. *Development*, 121(2):273–85, 1995.

G. Felsenfeld. Chromatin as an essential part of the transcriptional mechanism. *Nature*, 355 (6357):219–24, 1992.

Y. Q. Feng, J. Seibler, R. Alami, A. Eisen, K. A. Westerman, P. Leboulch, S. Fiering, and E. E. Bouhassira. Site-specific chromosomal integration in mammalian cells: highly efficient cre recombinase-mediated cassette exchange. *J Mol Biol*, 292(4):779–85, 1999.

W. Fischle, Y. Wang, S. A. Jacobs, Y. Kim, C. D. Allis, and S. Khorasanizadeh. Molecular basis for the discrimination of repressive methyl-lysine marks in histone h3 by polycomb and hp1 chromodomains. *Genes Dev*, 17(15):1870–81, 2003.

W. Fischle, B. S. Tseng, H. L. Dormann, B. M. Ueberheide, B. A. Garcia, J. Shabanowitz, D. F. Hunt, H. Funabiki, and C. D. Allis. Regulation of hp1-chromatin binding by histone h3 methylation and phosphorylation. *Nature*, 438(7071):1116–22, 2005.

N. J. Francis, R. E. Kingston, and C. L. Woodcock. Chromatin compaction by a polycomb group protein complex. *Science*, 306(5701):1574–7, 2004.

N. Frankel, D. F. Erezyilmaz, A. P. McGregor, S. Wang, F. Payre, and D. L. Stern. Morphological evolution caused by many subtle-effect substitutions in regulatory dna. *Nature*, 474(7353):598–603, 2011.

F. Frederiks, M. Tzouros, G. Oudgenoeg, T. van Welsem, M. Fornerod, J. Krijgsveld, and F. van Leeuwen. Nonprocessive methylation by dot1 leads to functional redundancy of histone h3k79 methylation states. *Nat Struct Mol Biol*, 15(6):550–7, 2008.

D. J. Gaffney, J. B. Veyrieras, J. F. Degner, R. Pique-Regi, A. A. Pai, G. E. Crawford, M. Stephens, Y. Gilad, and J. K. Pritchard. Dissecting the regulatory architecture of gene expression qtls. *Genome Biol*, 13(1):R7, 2012.

F. Gao, B. C. Foat, and H. J. Bussemaker. Defining transcriptional networks through integrative modeling of mrna expression and transcription factor binding data. *BMC Bioinformatics*, 5:31, 2004.

Z. Gao, J. Zhang, R. Bonasio, F. Strino, A. Sawai, F. Parisi, Y. Kluger, and D. Reinberg.

Pcgf homologs, cbx proteins, and rybp define functionally distinct prc1 family complexes. *Mol Cell*, 45(3):344–56, 2012.

J. Garcia-Fernandez. The genesis and evolution of homeobox gene clusters. *Nat Rev Genet*, 6(12):881–92, 2005.

S. S. Gehani, S. Agrawal-Singh, N. Dietrich, N. S. Christophersen, K. Helin, and K. Hansen. Polycomb group protein displacement and gene activation through msk-dependent h3k27me3s28 phosphorylation. *Mol Cell*, 39(6):886–900, 2010.

R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004.

Y. Gilad, S. A. Rifkin, and J. K. Pritchard. Revealing the architecture of gene regulation: the promise of eqtl studies. *Trends Genet*, 24(8):408–15, 2008.

E. L. Greer and Y. Shi. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat Rev Genet*, 13(5):343–57, 2012.

T.R. Gregory. Animal genome size database. 2012.

S. I. Grewal and S. C. Elgin. Heterochromatin: new possibilities for the inheritance of structure. *Curr Opin Genet Dev*, 12(2):178–87, 2002.

M. Grunstein. Histone acetylation in chromatin structure and transcription. *Nature*, 389 (6649):349–52, 1997.

E. Guccione, F. Martinato, G. Finocchiaro, L. Luzi, L. Tizzoni, V. Dall' Olio, G. Zardo, C. Nervi, L. Bernard, and B. Amati. Myc-binding-site recognition in the human genome is determined by chromatin context. *Nat Cell Biol*, 8(7):764–70, 2006.

R. A. Gupta, N. Shah, K. C. Wang, J. Kim, H. M. Horlings, D. J. Wong, M. C. Tsai, T. Hung, P. Argani, and H. Y. Chang. Long non-coding rna hotair reprograms chromatin state to promote cancer metastasis. *Nature*, 464(7291):1071–6, 2010.

K. H. Hansen, A. P. Bracken, D. Pasini, N. Dietrich, S. S. Gehani, A. Monrad, J. Rappsilber, M. Lerdrup, and K. Helin. A model for transmission of the h3k27me3 epigenetic mark. *Nat Cell Biol*, 10(11):1291–300, 2008.

M. Harbers and P. Carninci. Tag-based approaches for transcriptome research and genome annotation. *Nat Methods*, 2(7):495–502, 2005.

C. A. Hassig, T. C. Fleischer, A. N. Billin, S. L. Schreiber, and D. E. Ayer. Histone deacetylase activity is required for full transcriptional repression by msin3a. *Cell*, 89(3):341–7, 1997.

R. D. Hawkins, G. C. Hon, C. Yang, J. E. Antosiewicz-Bourget, L. K. Lee, Q. M. Ngo, S. Klugman, K. A. Ching, L. E. Edsall, and B. Ren. Dynamic chromatin states in human es cells reveal potential regulatory sequences and genes involved in pluripotency. *Cell Res*, 21(10):1393–409, 2011.

Q. He, A. F. Bardet, B. Patton, J. Purvis, J. Johnston, A. Paulson, M. Gogol, A. Stark, and J. Zeitlinger. High conservation of transcription factor binding and evidence for combinatorial regulation across six drosophila species. *Nat Genet*, 43(5):414–20, 2011.

N. Heins, P. Malatesta, F. Cecconi, M. Nakafuku, K. L. Tucker, M. A. Hack, P. Chapouton, Y. A. Barde, and M. Gotz. Glial cells generate neurons: the role of the transcription factor pax6. *Nat Neurosci*, 5(4):308–15, 2002.

N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, and B. Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*, 39(3):311–8, 2007.

N. D. Heintzman, G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. Ye, L. K. Lee, R. K. Stuart, and B. Ren. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–12, 2009.

E. Heitz. Das heterochromatin der moose. *Jahrb Wiss Bot*, (69):762–818, 1928.

K. Hemavathy, S. I. Ashraf, and Y. T. Ip. Snail/slug family of repressors: slowly going into the fast lane of development and cancer. *Gene*, 257(1):1–12, 2000.

M. Hemberg and G. Kreiman. Conservation of transcription factor binding events predicts gene expression across species. *Nucleic Acids Res*, 39(16):7092–102, 2011.

Y. Hirabayashi and Y. Gotoh. Epigenetic control of neural precursor cell fate during development. *Nat Rev Neurosci*, 11(6):377–88, 2010.

E. Hodges, A. Molaro, C. O. Dos Santos, P. Thekkat, Q. Song, P. J. Uren, J. Park, J. Butler, and G. J. Hannon. Directional dna methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol Cell*, 44(1): 17–28, 2011.

B. Horard, C. Tatout, S. Poux, and V. Pirrotta. Structure of a polycomb response element

and in vitro binding of polycomb group complexes containing gaga factor. *Mol Cell Biol*, 20(9):3187–97, 2000.

Y. Huang, S. J. Myers, and R. Dingledine. Transcriptional repression by rest: recruitment of sin3a and histone deacetylase to neuronal genes. *Nat Neurosci*, 2(10):867–72, 1999.

D. J. Huebert, P. F. Kuan, S. Keles, and A. P. Gasch. Dynamic changes in nucleosome occupancy are not predictive of gene expression dynamics but linked to transcription and chromatin regulators. *Mol Cell Biol*, 2012.

J. Hunkapiller, Y. Shen, A. Diaz, G. Cagney, D. McCleary, M. Ramalho-Santos, N. Krogan, B. Ren, J. S. Song, and J. F. Reiter. Polycomb-like 3 promotes polycomb repressive complex 2 binding to cpg islands and embryonic stem cell self-renewal. *PLoS Genet*, 8(3):e1002576, 2012.

T. Irie, S. J. Park, R. Yamashita, M. Seki, T. Yada, S. Sugano, K. Nakai, and Y. Suzuki. Predicting promoter activities of primary human dna sequences. *Nucleic Acids Res*, 39 (11):e75, 2011.

K. Jepsen, D. Solum, T. Zhou, R. J. McEvilly, H. J. Kim, C. K. Glass, O. Hermanson, and M. G. Rosenfeld. Smrt-mediated repression of an h3k27 demethylase in progression from neural stem cell to neuron. *Nature*, 450(7168):415–9, 2007.

C. Jiang and B. F. Pugh. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet*, 10(3):161–72, 2009.

H. Jiang, A. Shukla, X. Wang, W. Y. Chen, B. E. Bernstein, and R. G. Roeder. Role for dpy-30 in es cell-fate specification by regulation of h3k4 methylation within bivalent domains. *Cell*, 144(4):513–25, 2011.

R. Jiang, Y. Lan, C. R. Norton, J. P. Sundberg, and T. Gridley. The slug gene is not essential for mesoderm or neural crest development in mice. *Dev Biol*, 198(2):277–85, 1998.

S. John, P. J. Sabo, R. E. Thurman, M. H. Sung, S. C. Biddie, T. A. Johnson, G. L. Hager, and J. A. Stamatoyannopoulos. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet*, 43(3):264–8, 2011.

D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–502, 2007.

R. Johnson, J. Samuel, C. K. Ng, R. Jauch, L. W. Stanton, and I. C. Wood. Evolution of the

vertebrate gene regulatory network controlled by the transcriptional repressor rest. *Mol Biol Evol*, 26(7):1491–507, 2009.

F. S. Jones and R. Meech. Knockout of rest/nrsf shows that the protein is a potent repressor of neuronally expressed genes in non-neural tissues. *Bioessays*, 21(5):372–6, 1999.

H. F. Jorgensen, Z. F. Chen, M. Merkenschlager, and A. G. Fisher. Is rest required for esc pluripotency? *Nature*, 457(7233):E4–5; discussion E7, 2009.

A. Kanhere, K. Viiri, C. C. Araujo, J. Rasaiyaah, R. D. Bouwman, W. A. Whyte, C. F. Pereira, E. Brookes, K. Walker, and R. G. Jenner. Short rnas are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Mol Cell*, 38(5): 675–88, 2010.

H. Kataoka, T. Murayama, M. Yokode, S. Mori, H. Sano, H. Ozaki, Y. Yokota, S. Nishikawa, and T. Kita. A novel snail-related transcription factor smuc regulates basic helix-loop-helix transcription factor activities via specific e-box motifs. *Nucleic Acids Res*, 28(2):626–33, 2000.

P. S. Kayne, U. J. Kim, M. Han, J. R. Mullen, F. Yoshizaki, and M. Grunstein. Extremely conserved histone h4 n terminus is dispensable for growth but essential for repressing the silent mating loci in yeast. *Cell*, 55(1):27–39, 1988.

M. C. Keogh, S. K. Kurdistani, S. A. Morris, S. H. Ahn, V. Podolny, S. R. Collins, M. Schuldiner, K. Chin, T. Punna, and N. J. Krogan. Cotranscriptional set2 methylation of histone h3 lysine 36 recruits a repressive rpd3 complex. *Cell*, 123(4):593–605, 2005.

P. Kerner, J. Hung, J. Behague, M. Le Gouar, G. Balavoine, and M. Vervoort. Insights into the evolution of the snail superfamily from metazoan wide molecular phylogenies and expression data in annelids. *BMC Evol Biol*, 9:94, 2009.

H. Kim, K. Kang, and J. Kim. Aebp2 as a potential targeting protein for polycomb repression complex prc2. *Nucleic Acids Res*, 37(9):2940–50, 2009.

T. G. Kim, J. C. Kraus, J. Chen, and Y. Lee. Jumonji, a critical factor for cardiac development, functions as a transcriptional repressor. *J Biol Chem*, 278(43):42247–55, 2003.

J. A. Knezetic and D. S. Luse. The presence of nucleosomes on a dna template prevents initiation by rna polymerase ii in vitro. *Cell*, 45(1):95–104, 1986.

C. M. Koch, R. M. Andrews, P. Flicek, S. C. Dillon, U. Karaoz, G. K. Clelland, S. Wilcox,

D. M. Beare, J. C. Fowler, and I. Dunham. The landscape of histone modifications across 1in five human cell lines. *Genome Res*, 17(6):691–707, 2007.

T. Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007.

N. J. Krogan, M. Kim, A. Tong, A. Golshani, G. Cagney, V. Canadien, D. P. Richards, B. K. Beattie, A. Emili, and J. Greenblatt. Methylation of histone h3 by set2 in saccharomyces cerevisiae is linked to transcriptional elongation by rna polymerase ii. *Mol Cell Biol*, 23 (12):4207–18, 2003.

M. Ku, R. P. Koche, E. Rheinbay, E. M. Mendenhall, M. Endoh, T. S. Mikkelsen, A. Presser, C. Nusbaum, X. Xie, and B. E. Bernstein. Genomewide analysis of prc1 and prc2 occupancy identifies two classes of bivalent domains. *PLoS Genet*, 4(10):e1000242, 2008.

M. H. Kuo, J. Zhou, P. Jambeck, M. E. Churchill, and C. D. Allis. Histone acetyltransferase activity of yeast gcn5p is required for the activation of target genes in vivo. *Genes Dev*, 12(5):627–39, 1998.

M. Lachner, D. O'Carroll, S. Rea, K. Mechtler, and T. Jenuwein. Methylation of histone h3 lysine 9 creates a binding site for hp1 proteins. *Nature*, 410(6824):116–20, 2001.

D. Landeira, S. Sauer, R. Poot, M. Dvorkina, L. Mazzarella, H. F. Jorgensen, C. F. Pereira, M. Leleu, F. M. Piccolo, and A. G. Fisher. Jarid2 is a prc2 component in embryonic stem cells required for multi-lineage differentiation and recruitment of prc1 and rna polymerase ii to developmental regulators. *Nat Cell Biol*, 12(6):618–24, 2010.

E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, and Y. J. Chen. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

C. Lanzuolo, V. Roure, J. Dekker, F. Bantignies, and V. Orlando. Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex. *Nat Cell Biol*, 9(10):1167–74, 2007.

J. A. Law and S. E. Jacobsen. Establishing, maintaining and modifying dna methylation patterns in plants and animals. *Nat Rev Genet*, 11(3):204–20, 2010.

J. H. Lee and D. G. Skalnik. Cpg-binding protein (cxxc finger protein 1) is a component of

the mammalian set1 histone h3-lys4 methyltransferase complex, the analogue of the yeast set1/compass complex. *J Biol Chem*, 280(50):41725–31, 2005.

M. G. Lee, C. Wynder, N. Cooch, and R. Shiekhattar. An essential role for corest in nucleosomal histone 3 lysine 4 demethylation. *Nature*, 437(7057):432–5, 2005.

S. Lee, J. W. Lee, and S. K. Lee. Utx, a histone h3-lysine 27 demethylase, acts as a critical switch to activate the cardiac developmental program. *Dev Cell*, 22(1):25–37, 2011.

T. I. Lee, R. G. Jenner, L. A. Boyer, M. G. Guenther, S. S. Levine, R. M. Kumar, B. Chevalier, S. E. Johnstone, M. F. Cole, and R. A. Young. Control of developmental regulators by polycomb in human embryonic stem cells. *Cell*, 125(2):301–13, 2006.

T. L. Lenstra, J. J. Benschop, T. Kim, J. M. Schulze, N. A. Brabers, T. Margaritis, L. A. van de Pasch, S. A. van Heesch, M. O. Brok, and F. C. Holstege. The specificity and topology of chromatin interaction pathways in yeast. *Mol Cell*, 42(4):536–49, 2011.

M. Levine and R. Tjian. Transcription regulation and animal diversity. *Nature*, 424(6945): 147–51, 2003.

E. B. Lewis. A gene complex controlling segmentation in drosophila. *Nature*, 276(5688): 565–70, 1978.

B. Li, M. Carey, and J. L. Workman. The role of chromatin during transcription. *Cell*, 128 (4):707–19, 2007.

E. Li, T. H. Bestor, and R. Jaenisch. Targeted mutation of the dna methyltransferase gene results in embryonic lethality. *Cell*, 69(6):915–26, 1992.

G. Li, R. Margueron, M. Ku, P. Chambon, B. E. Bernstein, and D. Reinberg. Jarid2 and prc2, partners in regulating gene expression. *Genes Dev*, 24(4):368–80, 2010.

F. Lienert, F. Mohn, V. K. Tiwari, T. Baubec, T. C. Roloff, D. Gaidatzis, M. B. Stadler, and D. Schubeler. Genomic prevalence of heterochromatic h3k9me2 and transcription do not discriminate pluripotent from terminally differentiated cells. *PLoS Genet*, 7(6):e1002090, 2011a.

F. Lienert, C. Wirbelauer, I. Som, A. Dean, F. Mohn, and D. Schubeler. Identification of genetic elements that autonomously determine dna methylation states. *Nat Genet*, 43(11): 1091–7, 2011b.

S. Lisser and H. Margalit. Compilation of e. coli mrna promoter sequences. *Nucleic Acids Res*, 21(7):1507–16, 1993.

C. L. Liu, T. Kaplan, M. Kim, S. Buratowski, S. L. Schreiber, N. Friedman, and O. J. Rando. Single-nucleosome mapping of histone modifications in s. cerevisiae. *PLoS Biol*, 3(10):e328, 2005.

Y. Liu, S. El-Naggar, D. S. Darling, Y. Higashi, and D. C. Dean. Zeb1 links epithelial-mesenchymal transition and cellular senescence. *Development*, 135(3):579–88, 2008.

Y. Liu, Z. Shao, and G. C. Yuan. Prediction of polycomb target genes in mouse embryonic stem cells. *Genomics*, 96(1):17–26, 2010.

Y. Lorch, J. W. LaPointe, and R. D. Kornberg. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell*, 49(2):203–10, 1987.

X. Lu, M. D. Simon, J. V. Chodaparambil, J. C. Hansen, K. M. Shokat, and K. Luger. The effect of h3k79 dimethylation and h4k20 trimethylation on nucleosome and chromatin structure. *Nat Struct Mol Biol*, 15(10):1122–4, 2008.

R. F. Luco, Q. Pan, K. Tominaga, B. J. Blencowe, O. M. Pereira-Smith, and T. Misteli. Regulation of alternative splicing by histone modifications. *Science*, 327(5968):996–1000, 2010.

K. Luger, A. W. Mader, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, 389(6648):251–60, 1997.

M. D. Lynch, A. J. Smith, M. De Gobbi, M. Flenley, J. R. Hughes, D. Vernimmen, H. Ayyub, J. A. Sharpe, J. A. Sloane-Stanley, and D. R. Higgs. An interspecies analysis reveals a key role for unmethylated cpg dinucleotides in vertebrate polycomb complex recruitment. *EMBO J*, 31(2):317–29, 2011.

D. Macleod, J. Charlton, J. Mullins, and A. P. Bird. Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the cpg island. *Genes Dev*, 8(19):2282–92, 1994.

S. Majumder. Rest in good times and bad: roles in tumor suppressor and oncogenic activities. *Cell Cycle*, 5(17):1929–35, 2006.

M. Manzanares, A. Locascio, and M. A. Nieto. The increasing complexity of the snail gene superfamily in metazoan evolution. *Trends Genet*, 17(4):178–81, 2001.

R. Margueron and D. Reinberg. Chromatin structure and the inheritance of epigenetic information. *Nat Rev Genet*, 11(4):285–96, 2010.

R. Margueron, N. Justin, K. Ohno, M. L. Sharpe, J. Son, 3rd Drury, W. J., P. Voigt, S. R. Martin, W. R. Taylor, and S. J. Gamblin. Role of the polycomb protein eed in the propagation of repressive histone marks. *Nature*, 461(7265):762–7, 2009.

V. Mauhin, Y. Lutz, C. Dennefeld, and A. Alberga. Definition of the dna-binding site repertoire for the drosophila transcription factor snail. *Nucleic Acids Res*, 21(17):3951–7, 1993.

O. G. McDonald, H. Wu, W. Timp, A. Doi, and A. P. Feinberg. Genome-scale epigenetic reprogramming during epithelial-to-mesenchymal transition. *Nat Struct Mol Biol*, 18(8): 867–74, 2011.

A. Meissner. Epigenetic modifications in pluripotent and differentiated cells. *Nat Biotechnol*, 28(10):1079–88, 2010.

E. M. Mendenhall, R. P. Koche, T. Truong, V. W. Zhou, B. Issac, A. S. Chi, M. Ku, and B. E. Bernstein. Gc-rich sequence elements recruit prc2 in mammalian es cells. *PLoS Genet*, 6(12):e1001244, 2010.

J. R. Menezes and M. B. Luskin. Expression of neuron-specific tubulin defines a novel population in the proliferative layers of the developing telencephalon. *J Neurosci*, 14(9): 5399–416, 1994.

T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T. K. Kim, and B. E. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–60, 2007.

T. S. Mikkelsen, Z. Xu, X. Zhang, L. Wang, J. M. Gimble, E. S. Lander, and E. D. Rosen. Comparative epigenomic analysis of murine and human adipogenesis. *Cell*, 143(1):156–69, 2010.

A. Mohd-Sarip, A. Lagarou, C. M. Doyen, J. A. van der Knaap, U. Aslan, K. Bezstarosti, Y. Yassin, H. W. Brock, J. A. Demmers, and C. P. Verrijzer. Transcription-independent function of polycomb group protein psc in cell cycle control. *Science*, 2012.

F. Mohn and D. Schubeler. Genetics and epigenetics: stability and plasticity during cellular differentiation. *Trends Genet*, 25(3):129–36, 2009.

F. Mohn, M. Weber, M. Rebhan, T. C. Roloff, J. Richter, M. B. Stadler, M. Bibel, and

D. Schubeler. Lineage-specific polycomb targets and de novo dna methylation define restriction and potential of neuronal progenitors. *Mol Cell*, 30(6):755–66, 2008.

A. C. Mullen, D. A. Orlando, J. J. Newman, J. Loven, R. M. Kumar, S. Bilodeau, J. Reddy, M. G. Guenther, R. P. DeKoter, and R. A. Young. Master transcription factors determine cell-type-specific responses to tgf-beta signaling. *Cell*, 147(3):565–76, 2011.

N. V. Murzina, X. Y. Pei, W. Zhang, M. Sparkes, J. Vicente-Garcia, J. V. Pratap, S. H. McLaughlin, T. R. Ben-Shahar, A. Verreault, and E. D. Laue. Structural basis for the recognition of histone h4 by the histone-chaperone rbap46. *Structure*, 16(7):1077–85, 2008.

E. K. Nakakura, D. N. Watkins, K. E. Schuebel, V. Sriuranpong, M. W. Borges, B. D. Nelkin, and D. W. Ball. Mammalian scratch: a neural-specific snail family transcriptional repressor. *Proc Natl Acad Sci U S A*, 98(7):4010–5, 2001a.

E. K. Nakakura, D. N. Watkins, V. Sriuranpong, M. W. Borges, B. D. Nelkin, and D. W. Ball. Mammalian scratch participates in neuronal differentiation in p19 embryonal carcinoma cells. *Brain Res Mol Brain Res*, 95(1-2):162–6, 2001b.

S. Nakanishi, B. W. Sanderson, K. M. Delventhal, W. D. Bradford, K. Staehling-Hampton, and A. Shilatifard. A comprehensive library of histone mutants identifies nucleosomal residues required for h3k4 methylation. *Nat Struct Mol Biol*, 15(8):881–8, 2008.

H. Nakayama, I. C. Scott, and J. C. Cross. The transition to endoreduplication in trophoblast giant cells is regulated by the msna zinc finger transcription factor. *Dev Biol*, 199(1):150–63, 1998.

J. Nakayama, J. C. Rice, B. D. Strahl, C. D. Allis, and S. I. Grewal. Role of histone h3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science*, 292(5514):110–3, 2001.

X. Nan, H. H. Ng, C. A. Johnson, C. D. Laherty, B. M. Turner, R. N. Eisenman, and A. Bird. Transcriptional repression by the methyl-cpg-binding protein mecp2 involves a histone deacetylase complex. *Nature*, 393(6683):386–9, 1998.

Y. Nibu, H. Zhang, E. Bajor, S. Barolo, S. Small, and M. Levine. dctbp mediates transcriptional repression by knirps, kruppel and snail in the drosophila embryo. *EMBO J*, 17(23): 7009–20, 1998.

M. A. Nieto. The snail superfamily of zinc-finger transcription factors. *Nat Rev Mol Cell Biol*, 3(3):155–66, 2002.

M. A. Nieto, M. F. Bennett, M. G. Sargent, and D. G. Wilkinson. Cloning and developmental expression of sna, a murine homologue of the drosophila snail gene. *Development*, 116(1): 227–37, 1992.

J. P. Noonan. Regulatory dnas and the evolution of human development. *Curr Opin Genet Dev*, 19(6):557–64, 2009.

D. Noordermeer, M. Leleu, E. Splinter, J. Rougemont, W. De Laat, and D. Duboule. The dynamic architecture of hox gene clusters. *Science*, 334(6053):222–5, 2011.

D. O'Carroll, S. Erhardt, M. Pagani, S. C. Barton, M. A. Surani, and T. Jenuwein. The polycomb-group gene ezh2 is required for early mouse development. *Mol Cell Biol*, 21(13): 4330–6, 2001.

H. O'Geen, S. L. Squazzo, S. Iyengar, K. Blahnik, J. L. Rinn, H. Y. Chang, R. Green, and P. J. Farnham. Genome-wide analysis of kap1 binding suggests autoregulation of krab-znfs. *PLoS Genet*, 3(6):e89, 2007.

V. V. Ogryzko, R. L. Schiltz, V. Russanova, B. H. Howard, and Y. Nakatani. The transcriptional coactivators p300 and cbp are histone acetyltransferases. *Cell*, 87(5):953–9, 1996.

S. K. Ooi, C. Qiu, E. Bernstein, K. Li, D. Jia, Z. Yang, H. Erdjument-Bromage, P. Tempst, S. P. Lin, and T. H. Bestor. Dnmt3l connects unmethylated lysine 4 of histone h3 to de novo methylation of dna. *Nature*, 448(7154):714–7, 2007.

K. Palm, N. Belluardo, M. Metsis, and T. Timmusk. Neuronal expression of zinc finger transcription factor rest/nrsf/xbr gene. *J Neurosci*, 18(4):1280–96, 1998.

G. Pan, S. Tian, J. Nie, C. Yang, V. Ruotti, H. Wei, G. A. Jonsdottir, R. Stewart, and J. A. Thomson. Whole-genome analysis of histone h3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell*, 1(3):299–312, 2007.

A. J. Paquette, S. E. Perez, and D. J. Anderson. Constitutive expression of the neuron-restrictive silencer factor (nrsf)/rest in differentiating neurons disrupts neuronal gene expression and causes axon pathfinding errors in vivo. *Proc Natl Acad Sci U S A*, 97(22): 12318–23, 2000.

T. J. Parnell, J. T. Huff, and B. R. Cairns. Rsc regulates nucleosome positioning at pol ii genes and density at pol iii genes. *EMBO J*, 27(1):100–10, 2008.

D. Pasini, A. P. Bracken, M. R. Jensen, E. Lazzerini Denchi, and K. Helin. Suz12 is essential

for mouse development and for ezh2 histone methyltransferase activity. *EMBO J*, 23(20): 4061–71, 2004.

D. Pasini, P. A. Cloos, J. Walfridsson, L. Olsson, J. P. Bukowski, J. V. Johansen, M. Bak, N. Tommerup, J. Rappsilber, and K. Helin. Jarid2 regulates binding of the polycomb repressive complex 2 to target genes in es cells. *Nature*, 464(7286):306–10, 2010.

M. Paulsen and A. C. Ferguson-Smith. Dna methylation in genomic imprinting, development, and disease. *J Pathol*, 195(1):97–110, 2001.

H. Peinado, E. Ballestar, M. Esteller, and A. Cano. Snail mediates e-cadherin repression by the recruitment of the sin3a/histone deacetylase 1 (hdac1)/hdac2 complex. *Mol Cell Biol*, 24(1):306–19, 2004.

S. Peiro, M. Escriva, I. Puig, M. J. Barbera, N. Dave, N. Herranz, M. J. Larriba, M. Takkunen, C. Franci, and A. Garcia de Herreros. Snail1 transcriptional repressor binds to its own promoter and controls its expression. *Nucleic Acids Res*, 34(7):2077–84, 2006.

J. C. Peng, A. Valouev, T. Swigut, J. Zhang, Y. Zhao, A. Sidow, and J. Wysocka. Jarid2/jumonji coordinates control of prc2 enzymatic activity and target gene occupancy in pluripotent cells. *Cell*, 139(7):1290–302, 2009.

C. F. Pereira, F. M. Piccolo, T. Tsubouchi, S. Sauer, N. K. Ryan, L. Bruno, D. Landeira, J. Santos, A. Banito, and A. G. Fisher. Escs require prc2 to direct the successful reprogramming of differentiated cells toward pluripotency. *Cell Stem Cell*, 6(6):547–56, 2010.

R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard. Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data. *Genome Res*, 21(3):447–55, 2011.

N. Plachta, M. Bibel, K. L. Tucker, and Y. A. Barde. Developmental potential of defined neural progenitors derived from mouse embryonic stem cells. *Development*, 131(21):5449–56, 2004.

D. K. Pokholok, C. T. Harbison, S. Levine, M. Cole, N. M. Hannett, T. I. Lee, G. W. Bell, K. Walker, P. A. Rolfe, and R. A. Young. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, 122(4):517–27, 2005.

A. A. Postigo and D. C. Dean. Differential expression and function of members of the zfh-1 family of zinc finger/homeodomain repressors. *Proc Natl Acad Sci U S A*, 97(12):6391–6, 2000.

M. G. Pray-Grant, J. A. Daniel, D. Schieltz, 3rd Yates, J. R., and P. A. Grant. Chd1 chromodomain links histone h3 methylation with saga- and slik-dependent acetylation. *Nature*, 433(7024):434–8, 2005.

M. Ptashne and A. Gann. *Genes and Signals*. Cold Spring Harbor, New York, 2002.

E. Pujadas and A. P. Feinberg. Regulated noise in the epigenetic landscape of development and disease. *Cell*, 148(6):1123–31, 2012.

O. J. Rando. Global patterns of histone modifications. *Curr Opin Genet Dev*, 17(2):94–9, 2007.

O. J. Rando. Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Curr Opin Genet Dev*, 22(2):148–55, 2012.

X. Ren and T. K. Kerppola. Rest interacts with cbx proteins and regulates polycomb repressive complex 1 occupancy at re1 elements. *Mol Cell Biol*, 31(10):2100–10, 2011.

N. Reynolds, M. Salmon-Divon, H. Dvinge, A. Hynes-Allen, G. Balasooriya, D. Leaford, A. Behrens, P. Bertone, and B. Hendrich. Nurd-mediated deacetylation of h3k27 facilitates recruitment of polycomb repressive complex 2 to direct gene repression. *EMBO J*, 31(3): 593–605, 2011.

H. Richly, L. Aloia, and L. Di Croce. Roles of the polycomb group proteins in stem cells and cancer. *Cell Death Dis*, 2:e204, 2011.

L. Ringrose and R. Paro. Epigenetic regulation of cellular memory by the polycomb and trithorax group proteins. *Annu Rev Genet*, 38:413–43, 2004.

L. Ringrose and R. Paro. Polycomb/trithorax response elements and epigenetic memory of cell identity. *Development*, 134(2):223–32, 2007.

J. L. Rinn, M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, J. A. Helms, P. J. Farnham, and H. Y. Chang. Functional demarcation of active and silent chromatin domains in human hox loci by noncoding rnas. *Cell*, 129(7):1311–23, 2007.

G. Ruvkun and O. Hobert. The taxonomy of developmental control in caenorhabditis elegans. *Science*, 282(5396):2033–41, 1998.

H. Santos-Rosa, R. Schneider, A. J. Bannister, J. Sherriff, B. E. Bernstein, N. C. Emre, S. L. Schreiber, J. Mellor, and T. Kouzarides. Active genes are tri-methylated at k4 of histone h3. *Nature*, 419(6905):407–11, 2002.

J. S. Satterlee, D. Schubeler, and H. H. Ng. Tackling the epigenome: challenges and opportunities for collaboration. *Nat Biotechnol*, 28(10):1039–44, 2010.

F. W. Schmitges, A. B. Prusty, M. Faty, A. Stutzer, G. M. Lingaraju, J. Aiwazian, R. Sack, D. Hess, L. Li, and N. H. Thoma. Histone methylation by prc2 is inhibited by active chromatin marks. *Mol Cell*, 42(3):330–41, 2011.

S. Schoeftner, A. K. Sengupta, S. Kubicek, K. Mechtler, L. Spahn, H. Koseki, T. Jenuwein, and A. Wutz. Recruitment of prc1 function at the initiation of x inactivation independent of prc2 and silencing. *EMBO J*, 25(13):3110–22, 2006.

P. Schorderet and D. Duboule. Structural and functional differences in the long non-coding rna hotair in mouse and human. *PLoS Genet*, 7(5):e1002071, 2011.

G. Schotta, M. Lachner, K. Sarma, A. Ebert, R. Sengupta, G. Reuter, D. Reinberg, and T. Jenuwein. A silencing pathway to induce h3-k9 and h4-k20 trimethylation at constitutive heterochromatin. *Genes Dev*, 18(11):1251–62, 2004.

D. Schubeler, D. M. MacAlpine, D. Scalzo, C. Wirbelauer, C. Kooperberg, F. van Leeuwen, D. E. Gottschling, L. P. O'Neill, B. M. Turner, and M. Groudine. The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev*, 18(11):1263–71, 2004.

B. Schuettengruber and G. Cavalli. Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice. *Development*, 136(21):3531–42, 2009.

B. Schuettengruber, D. Chourrout, M. Vervoort, B. Leblanc, and G. Cavalli. Genome regulation by polycomb and trithorax proteins. *Cell*, 128(4):735–45, 2007.

S. Schwartz, E. Meshorer, and G. Ast. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol*, 16(9):990–5, 2009.

Y. B. Schwartz and V. Pirrotta. Polycomb silencing mechanisms and the management of genomic programmes. *Nat Rev Genet*, 8(1):9–22, 2007.

Y. B. Schwartz and V. Pirrotta. Polycomb complexes and epigenetic states. *Curr Opin Cell Biol*, 20(3):266–73, 2008.

M. Sefton, S. Sanchez, and M. A. Nieto. Conserved and divergent roles for members of the snail family of transcription factors in the chick and mouse embryo. *Development*, 125(16): 3111–21, 1998.

Y. Shi, J. Sawada, G. Sui, B. Affar el, J. R. Whetstine, F. Lan, H. Ogawa, M. P. Luke, and

Y. Nakatani. Coordinated histone modifications mediated by a ctbp co-repressor complex. *Nature*, 422(6933):735–8, 2003.

M. Shogren-Knaak, H. Ishii, J. M. Sun, M. J. Pazin, J. R. Davie, and C. L. Peterson. Histone h4-k16 acetylation controls chromatin structure and protein interactions. *Science*, 311(5762):844–7, 2006.

J. A. Simon and R. E. Kingston. Mechanisms of polycomb gene silencing: knowns and unknowns. *Nat Rev Mol Cell Biol*, 10(10):697–708, 2009.

3rd Sims, R. J., C. F. Chen, H. Santos-Rosa, T. Kouzarides, S. S. Patel, and D. Reinberg. Human but not yeast chd1 binds directly and selectively to histone h3 methylated at lysine 4 via its tandem chromodomains. *J Biol Chem*, 280(51):41789–92, 2005.

A. Sing, D. Pannell, A. Karaiskakis, K. Sturgeon, M. Djabali, J. Ellis, H. D. Lipshitz, and S. P. Cordes. A vertebrate polycomb response element governs segmentation of the posterior hindbrain. *Cell*, 138(5):885–97, 2009.

A. Sparmann and M. van Lohuizen. Polycomb silencers control cell fate, development and cancer. *Nat Rev Cancer*, 6(11):846–56, 2006.

S. L. Squazzo, H. O'Geen, V. M. Komashko, S. R. Krig, V. X. Jin, S. W. Jang, R. Margueron, D. Reinberg, R. Green, and P. J. Farnham. Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome Res*, 16(7):890–900, 2006.

M. B. Stadler, R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Scholer, C. Wirbelauer, E. J. Oakeley, D. Gaidatzis, and D. Schubeler. Dna-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480(7378):490–5, 2011.

J. K. Stock, S. Giadrossi, M. Casanova, E. Brookes, M. Vidal, H. Koseki, N. Brockdorff, A. G. Fisher, and A. Pombo. Ring1-mediated ubiquitination of h2a restrains poised rna polymerase ii at bivalent genes in mouse es cells. *Nat Cell Biol*, 9(12):1428–35, 2007.

B. D. Strahl and C. D. Allis. The language of covalent histone modifications. *Nature*, 403 (6765):41–5, 2000.

K. Struhl. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, 98(1):1–4, 1999.

Y. M. Sun, D. J. Greenway, R. Johnson, M. Street, N. D. Belyaev, J. Deuchars, T. Bee, S. Wilde, and N. J. Buckley. Distinct profiles of rest interactions with its target genes at different stages of neuronal development. *Mol Biol Cell*, 16(12):5630–8, 2005.

H. Suzuki, A. R. Forrest, E. van Nimwegen, C. O. Daub, P. J. Balwierz, K. M. Irvine, T. Lassmann, T. Ravasi, Y. Hasegawa, and Y. Hayashizaki. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet*, 41(5):553–62, 2009.

M. M. Suzuki and A. Bird. Dna methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet*, 9(6):465–76, 2008.

E. Szathmary, F. Jordan, and C. Pal. Molecular biology and evolution. can genes explain biological complexity? *Science*, 292(5520):1315–6, 2001.

Y. Tao, R. F. Kassatly, W. D. Cress, and J. M. Horowitz. Subunit composition determines e2f dna-binding site specificity. *Mol Cell Biol*, 17(12):6994–7007, 1997.

J. Taunton, C. A. Hassig, and S. L. Schreiber. A mammalian histone deacetylase related to the yeast transcriptional regulator rpd3p. *Science*, 272(5260):408–11, 1996.

L. Tavares, E. Dimitrova, D. Oxley, J. Webster, R. Poot, J. Demmers, K. Bezstarosti, S. Taylor, H. Ura, and N. Brockdorff. Rybp-prc1 complexes mediate h2a ubiquitylation at polycomb target sites independently of prc2 and h3k27me3. *Cell*, 148(4):664–78, 2012.

J. P. Thomson, P. J. Skene, J. Selfridge, T. Clouaire, J. Guy, S. Webb, A. R. Kerr, A. Deaton, R. Andrews, and A. Bird. Cpg islands influence chromatin structure via the cpg-binding protein cfp1. *Nature*, 464(7291):1082–6, 2010.

V. K. Tiwari, L. Cope, K. M. McGarvey, J. E. Ohm, and S. B. Baylin. A novel 6c assay uncovers polycomb-mediated higher order chromatin conformations. *Genome Res*, 18(7): 1171–9, 2008.

T. Tohyama, V. M. Lee, L. B. Rorke, M. Marvin, R. D. McKay, and J. Q. Trojanowski. Nestin expression in embryonic human neuroepithelium and in human neuroepithelial tumor cells. *Lab Invest*, 66(3):303–13, 1992.

B. Tolhuis, E. de Wit, I. Muijrers, H. Teunissen, W. Talhout, B. van Steensel, and M. van Lohuizen. Genome-wide profiling of prc1 and prc2 polycomb chromatin binding in drosophila melanogaster. *Nat Genet*, 38(6):694–9, 2006.

M. C. Tsai, O. Manor, Y. Wan, N. Mosammaparast, J. K. Wang, F. Lan, Y. Shi, E. Segal, and H. Y. Chang. Long noncoding rna as modular scaffold of histone modification complexes. *Science*, 329(5992):689–93, 2010.

R. Tupler, G. Perini, and M. R. Green. Expressing the human genome. *Nature*, 409(6822): 832–3, 2001.

M. Turner. Is transcription the dominant force during dynamic changes in gene expression? *Adv Exp Med Biol*, 780:1–13, 2011.

F. van Leeuwen, P. R. Gafken, and D. E. Gottschling. Dot1p modulates silencing in yeast by methylation of the nucleosome core. *Cell*, 109(6):745–56, 2002.

E. van Nimwegen. Scaling laws in the functional content of genomes. *Trends Genet*, 19(9): 479–84, 2003.

E. van Nimwegen. Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics*, 8 Suppl 6:S4, 2007.

J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, 10(4): 252–63, 2009.

S. Varambally, S. M. Dhanasekaran, M. Zhou, T. R. Barrette, C. Kumar-Sinha, M. G. Sanda, D. Ghosh, K. J. Pienta, R. G. Sewalt, and A. M. Chinnaiyan. The polycomb group protein ezh2 is involved in progression of prostate cancer. *Nature*, 419(6907):624–9, 2002.

A. Verdel, S. Jia, S. Gerber, T. Sugiyama, S. Gygi, S. I. Grewal, and D. Moazed. Rnai-mediated targeting of heterochromatin by the rits complex. *Science*, 303(5658):672–6, 2004.

A. Verreault, P. D. Kaufman, R. Kobayashi, and B. Stillman. Nucleosome assembly by a complex of caf-1 and acetylated histones h3/h4. *Cell*, 87(1):95–104, 1996.

M. K. Vickaryous and B. K. Hall. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol Rev Camb Philos Soc*, 81(3):425–55, 2006.

D. Vlieghe, A. Sandelin, P. J. De Bleser, K. Vleminckx, W. W. Wasserman, F. van Roy, and B. Lenhard. A new generation of jaspar, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res*, 34(Database issue):D95–7, 2006.

C.H. Waddington. The epigenotype. *Endeavour*, 1:1820, 1942.

C.H. Waddington. *The Strategy of the Genes; a Discussion of Some Aspects of Theoretical Biology*. Allen and Unwin, London, 1957.

H. Wang, L. Wang, H. Erdjument-Bromage, M. Vidal, P. Tempst, R. S. Jones, and Y. Zhang. Role of histone h2a ubiquitination in polycomb silencing. *Nature*, 431(7010):873–8, 2004.

J. Wang, J. Mager, Y. Chen, E. Schneider, J. C. Cross, A. Nagy, and T. Magnuson. Imprinted x inactivation maintained by a mouse polycomb group gene. *Nat Genet*, 28(4):371–5, 2001.

Z. Wang, C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, W. Peng, and K. Zhao. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet*, 40(7):897–903, 2008.

W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*, 5(4):276–87, 2004.

M. Weber, I. Hellmann, M. B. Stadler, L. Ramos, S. Paabo, M. Rebhan, and D. Schubeler. Distribution, silencing potential and evolutionary impact of promoter dna methylation in the human genome. *Nat Genet*, 39(4):457–66, 2007.

T. F. Westbrook, E. S. Martin, M. R. Schlabach, Y. Leng, A. C. Liang, B. Feng, J. J. Zhao, T. M. Roberts, G. Mandel, and S. J. Elledge. A genetic screen for candidate tumor suppressors identifies rest. *Cell*, 121(6):837–48, 2005.

I. Whitehouse, O. J. Rando, J. Delrow, and T. Tsukiyama. Chromatin remodelling at promoters suppresses antisense transcription. *Nature*, 450(7172):1031–5, 2007.

M. D. Wilson and D. T. Odom. Evolution of transcriptional control in mammals. *Curr Opin Genet Dev*, 19(6):579–85, 2009.

E. Wingender, P. Dietze, H. Karas, and R. Knuppel. Transfac: a database on transcription factors and their dna binding sites. *Nucleic Acids Res*, 24(1):238–41, 1996.

A. P. Wolffe and J. J. Hayes. Chromatin disruption and modification. *Nucleic Acids Res*, 27 (3):711–20, 1999.

C. J. Woo, P. V. Kharchenko, L. Daheron, P. J. Park, and R. E. Kingston. A region of the human hoxd cluster that confers polycomb-group responsiveness. *Cell*, 140(1):99–110, 2010.

J. L. Workman and R. E. Kingston. Nucleosome core displacement in vitro via a metastable transcription factor-nucleosome complex. *Science*, 258(5089):1780–4, 1992.

G. A. Wray, M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, and L. A. Romano. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol*, 20(9): 1377–419, 2003.

C. Wu, Y. C. Wong, and S. C. Elgin. The chromatin structure of specific genes: Ii. disruption of chromatin structure during gene activity. *Cell*, 16(4):807–14, 1979.

J. Wysocka, T. Swigut, H. Xiao, T. A. Milne, S. Y. Kwon, J. Landry, M. Kauer, A. J. Tackett, B. T. Chait, and C. D. Allis. A phd finger of nurf couples histone h3 lysine 4 trimethylation with chromatin remodelling. *Nature*, 442(7098):86–90, 2006.

Y. Yamada, H. Aoki, T. Kunisada, and A. Hara. Rest promotes the early differentiation of mouse escs but is not required for their maintenance. *Cell Stem Cell*, 6(1):10–5, 2010.

A. You, J. K. Tong, C. M. Grozinger, and S. L. Schreiber. Corest is an integral component of the corest- human histone deacetylase complex. *Proc Natl Acad Sci U S A*, 98(4):1454–8, 2001.

R. A. Young. Control of the embryonic stem cell state. *Cell*, 144(6):940–54, 2011.

M. Yu, T. Mazor, H. Huang, H. T. Huang, K. L. Kathrein, A. J. Woo, C. R. Chouinard, A. Labadorf, T. E. Akie, and A. B. Cantor. Direct recruitment of polycomb repressive complex 1 to chromatin by core binding transcription factors. *Mol Cell*, 45(3):330–43, 2012.

K. S. Zaret and J. S. Carroll. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev*, 25(21):2227–41, 2011.

J. Zeitlinger, R. P. Zinzen, A. Stark, M. Kellis, H. Zhang, R. A. Young, and M. Levine. Whole-genome chip-chip analysis of dorsal, twist, and snail suggests integration of diverse patterning processes in the drosophila embryo. *Genes Dev*, 21(4):385–90, 2007.

C. Zhang, Z. Xuan, S. Otto, J. R. Hover, S. R. McCorkle, G. Mandel, and M. Q. Zhang. A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. *Nucleic Acids Res*, 34(8):2238–46, 2006.

Z. Zhang, C. J. Wippo, M. Wal, E. Ward, P. Korber, and B. F. Pugh. A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science*, 332(6032): 977–80, 2011.

J. Zhao, B. K. Sun, J. A. Erwin, J. J. Song, and J. T. Lee. Polycomb proteins targeted by a short repeat rna to the mouse x chromosome. *Science*, 322(5902):750–6, 2008.

D. Zheng, K. Zhao, and M. F. Mehler. Profiling re1/rest-mediated histone modifications in the human genome. *Genome Biol*, 10(1):R9, 2009.

V. W. Zhou, A. Goren, and B. E. Bernstein. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet*, 12(1):7–18, 2011.